Identifying Unconditional Quantile Impulse Responses with an Application to Growth-at-Risk*

Robert Wojciechowski†

November 11, 2026
Latest Version

Abstract

I show that the causal effects of shocks on unconditional quantiles are identified with controls, and propose Generalized Quantile Local Projections (GQLP), a feasible estimator. Unlike quantile local projections, which estimate conditional quantile effects, GQLP estimates causal effects on unconditional quantiles while still exploiting controls for identification. This distinction is crucial in macroeconomics. For instance, when studying output growth, low unconditional quantiles correspond to actual recessions rather than merely periods of lower-than-expected growth relative to control variables. In a growth-at-risk application, I show that financial risk shocks have strongly asymmetric effects. Using timing restrictions for identification, I find that a one-standard-deviation increase in credit risk lowers the $\tau = 0.1$ quantile of cumulative industrial production growth by up to 1.8 percentage points, and the median by at most 0.84. Averaged over the first year, the response of the $\tau = 0.1$ quantile is 2.5 times that of the median for credit risk and 3.7 times for volatility risk, against 1.4 and 2.1 with conventional Quantile Local Projections, which estimate effects on conditional quantiles.

JEL Classification: C32, E44, C14, E32.

---

*Acknowledgments: I am indebted to Christian Brownlees and Andrea Caggese for their support and guidance. I would also like to thank Geert Mesters, Barbara Rossi, Luca Gambetti, Florens Odendahl, Vladislav Morozov, Valeria Gargiulo and the participants of the 48th Symposium of the Spanish Economic Association, the Euro Area Business Cycle Network: Advances in Local Projections and Empirical Methods for Central Banking conference, the 2024 European Winter Meeting of the Econometric Society, CREi Macroeconomics Lunch, the 2025 American Economic Association Annual Meeting, the Bank of England Financial Stability Seminar as well as seminar participants at the European Central Bank, Banco de España, Banco de Portugal, Universitat de les Illes Baleares, Universidad de Alicante, and Universitat Autònoma de Barcelona, for comments and discussion. I gratefully acknowledge support from the "Advances in Econometrics" grant (PID2022-140825NB-I00), funded by *Agencia Estatal de Investigación* (MCIN/AEI/10.13039/501100011033) and ESF+. All remaining errors are my own.

†Universitat Autònoma de Barcelona and Barcelona School of Economics, robert.wojciechowski@bse.eu

1 Introduction

Modern macroeconomics increasingly recognizes that shocks may affect the entire distribution of economic outcomes, beyond just the mean, and in particular the tails of the distribution. Explicit attention to tail risks has also become commonplace in policy. For instance, at a recent FOMC meeting (September 2025), the Committee motivated a federal funds rate cut by noting that "downside risks to employment have risen."

To understand the drivers of tail risks, researchers increasingly use quantile regression to measure how quantiles of the outcome distribution respond to shocks. However, a methodological complication arises: the inclusion of control variables for causal identification transforms the analysis from unconditional to conditional quantiles. This distinction is central across many areas of macroeconomics. When studying fluctuations in output growth or inflation, researchers are often interested in the drivers of extreme outcomes, such as recessions or destabilizing inflationary episodes. Unconditional tail quantiles directly correspond to these extreme outcome periods. In contrast, conditional tail quantiles correspond to periods that are extreme relative to what the control variables predict and thus do not always map onto actual crisis periods. For instance, a conditionally low quantile of output growth may correspond to a period of underperforming growth given favorable economic conditions, yet still occur during an expansion.

Quantile regression estimates coefficients using observations at specific quantiles of the conditional outcome distribution. Adding controls changes that conditional distribution, altering which observations are "local" to a given quantile and thereby shifting the estimated coefficients. In linear regression, adding controls orthogonal to the treatment leaves average treatment effects unchanged. However, in quantile regression even the addition of statistically independent controls can change the estimated quantile treatment effects. This creates a trade-off: unconditional quantiles without controls have clear economic interpretation but potentially biased estimates; including controls helps achieve causal identification but at the cost of shifting the analysis to conditional quantiles and thereby losing economic

interpretability.

This paper shows how to resolve this tension between causal identification and economically meaningful quantile interpretation. I show that the causal effect of a shock on the unconditional quantiles of an outcome, the quantile impulse response, is identified with controls. To estimate it, I introduce Generalized Quantile Local Projections (GQLP), a feasible estimator that builds on the generalized quantile regression of Powell (2020), the local projections approach of Jordá (2005), and the potential outcomes framework for time series of Rambachan and Shephard (2025). GQLP explicitly distinguishes treatment variables (whose effects we seek to measure) from control variables (used solely for identification), allowing estimation of causal effects on unconditional quantiles while still exploiting controls for identification. Beyond disentangling the effect of including controls on identification versus interpretation, GQLP has three additional advantages. First, the framework defines quantile impulse responses (QIRs) as responses to one-off shocks, ensuring direct comparability with conventional mean impulse responses. Second, it does not require the structural quantile function to be linear in the treatment[1], accommodating complex nonlinear responses that depend on both baseline and counterfactual treatment values. Third, while this paper emphasizes identification through timing restrictions and control variables, the framework naturally extends to instrumental variable designs.

Growth-at-risk (GaR) provides a natural application for GQLP. GaR is defined as the $\tau = 0.05$ quantile of future growth, representing a pessimistic scenario that materializes five percent of the time. A large empirical literature documents that financial conditions are key drivers of GaR (Adrian et al. 2019). This literature is primarily focused on forecasting (Plagborg-Møller et al. 2020; Brownlees and Souza 2021; Chuliá et al. 2024), with policymakers using declining GaR forecasts as early warnings of rising distress in the economy (Prasad et al. 2019). Theoretical macrofinance models attribute the heterogeneous effect of financial

[1]I focus on linear specifications in this paper, but nonlinear terms in the treatment can readily be added; appendix A2.3.3 estimates a quadratic and a spline in the treatment in the DSGE model, and appendix A3.3.1 a spline in the application.

conditions on upside versus downside risk to occasionally binding financial constraints, balance sheet interactions, and amplification mechanisms that can endogenously cause occasional financial crises (He and Krishnamurthy 2019; Gertler et al. 2019; Brunnermeier and Sannikov 2014).

I identify the causal effects of financial risk shocks on the unconditional quantiles of industrial production growth via timing restrictions, controlling for macroeconomic, financial, and monetary policy variables. The results indicate large and persistent output losses in low-growth environments, while effects at the median and upper quantiles are considerably smaller. For example, a one-standard-deviation credit risk shock reduces the $\tau = 0.1$ quantile of cumulative growth by up to 1.8 percentage points, compared to at most 0.84 percentage points at the median. Relative to conventional Quantile Local Projections (QLP), GQLP estimates uncover substantially stronger asymmetries between the lower tail and the median, with ratios of the lower-tail ($\tau = 0.1$) to the median response, averaged over the first year, of 2.5 for credit risk and 3.7 for volatility risk, compared to 1.4 and 2.1 under QLP. The difference arises mainly because QLP identifies effects on conditional quantiles, whereas GQLP recovers effects on unconditional quantiles, preserving the interpretation of quantiles as corresponding to phases of the business cycle. These results are in line with the fact that conditional quantile methods may underestimate adverse impacts of financial shocks during economic stress periods. I rationalize this finding in the Monte Carlo section with a stylized structural VAR in which an amplification mechanism is active only in bad states: conditioning on the state removes the amplification from the conditional lower tail, so that a quantile regression with controls understates the response of the lower tail.

To establish the theoretical foundation for quantile analysis, I prove a quantile invariance proposition. The proposition states that any dependent process with a structural Wold representation featuring independent innovations features quantile impulse responses that are identical across all conditional and unconditional quantiles and equal to the mean impulse response. This means that interesting quantile dynamics require either nonlinearity or

innovations that are uncorrelated but not independent. Since structural shocks are usually thought of as independent, in practice they require nonlinearity, and then a QLP that is linear in the treatment and the controls is generally misspecified. To compute the true structural quantile functions and QIRs in nonlinear models, I simulate the potential outcomes by fixing the treatment, at values from a grid for the structural quantile function and at randomly drawn values for the QIR, while drawing all other shocks, which follows directly from the definition of potential outcomes (appendix A2.1.2).

I do so in a nonlinear structural vector autoregression (SVAR) and a nonlinear dynamic stochastic general equilibrium (DSGE) model of Gertler et al. (2019). I show that in these models shocks have asymmetric effects across the distribution, with responses differing substantially between upper and lower quantiles. Moreover, the structural quantile functions exhibit nonlinear functional forms at some horizons for some quantiles. I then compare the performance of QLP with GQLP in a Monte Carlo study. Using the SVAR I show that GQLPs with timing restrictions produce estimates close to those of the infeasible regression on the structural shock, a property shared by standard SVARs and local projections but not by QLPs. This ensures GQLP-estimated QIRs identify the causal effect of a structural shock to the treatment on the quantile of the outcome. Using the DSGE model I show that GQLP works in a more realistic structural setting with occasionally binding constraints and bank runs, where its identifying assumptions need not hold exactly. The treatment is growth in the quality of bank assets, which responds to the lagged state of the economy. With controls for the lagged state, GQLP recovers a strongly asymmetric response of output in samples of realistic length. QLP without controls recovers only about a third of the response, and QLP with conventional linear controls misses the asymmetry on impact.

This work contributes to several strands of literature. I extend the generalized quantile regression of Powell (2020) from the cross-section quantile treatment effects literature (Chernozhukov and Hansen 2013; Angrist et al. 2006) to impulse response analysis. I also extend his asymptotic theory, which assumes independent observations, to stationary, weakly

dependent data. In doing so, I build on the potential outcomes for time series framework (Rambachan and Shephard 2025; Angrist and Kuersteiner 2011; Angrist et al. 2018). The methodology advances the local projections literature (Jordá 2005; Jorda and Taylor 2025) by enabling identification of unconditional quantile impulse responses. By establishing a novel causal link between financial risk shocks and the unconditional tail risks of output growth, I contribute to an emerging empirical literature that studies the structural drivers of GaR (Adrian et al. 2020; Adrian et al. 2022; Loria et al. 2025). My empirical findings can be rationalized by the macrofinance literature (Kiyotaki and Moore 1997; Brunnermeier and Sannikov 2014; He and Krishnamurthy 2019; Gertler et al. 2019) and are related to work on the cyclicality of volatility and the skewness of macroeconomic variables (see Bloom (2014) for a general review, and Cascaldi-Garcia et al. (2023) for a review of measures of uncertainty, volatility and risk).

Various alternative definitions and estimators of quantile impulse responses exist in the literature, but none addresses the fundamental challenge of identifying causal effects on unconditional quantiles while using control variables for identification. The QLP framework in Ruzicka (2026), Linnemann and Winkler (2016), Adrian et al. (2019), Adrian et al. (2022), Jordà et al. (2022), and Bochmann et al. (2023) is the closest to my approach. This framework recovers the QIR from local projection coefficients estimated using the quantile regression of Koenker and Bassett (1978), with estimation done separately for each quantile and horizon. In the absence of control variables, there is no difference between GQLP and QLP. Chavleishvili and Manganelli (2024) achieve identification by imposing timing restrictions on a recursive quantile vector autoregressive model. The reported QIRs assume a realization of a median sample path for the shock variable over the response horizon. Montes-Rojas (2019) and Lee et al. (2021) use a SVAR model to identify a structural shock since their multivariate quantile models are reduced-form. The QIR proposed by Montes-Rojas (2019) describes the cumulative impact of a series of shocks, not a one-off shock, because persistent realizations of lower (or upper) quantiles are assumed in its construction. Han et al. (2022) and Jung and

Lee (2022) study QIRs in models where the quantile itself is autoregressive, as in the CAViaR model of Engle and Manganelli (2004). In the applied literature, Mumtaz and Surico (2015) estimate structural QIRs of output growth in response to monetary policy shocks using the quantile autoregressive-distributed lag model of Galvao et al. (2013). Loria et al. (2025) estimate QIRs within a conventional local projections framework, but using as dependent variables the fitted quantiles of year-ahead output growth obtained from a quantile regression on current macro-financial conditions.

The remainder of the paper is structured as follows. Section 2 introduces the econometric framework. Section 3 motivates the framework using a simulation study. Section 4 contains the empirical analysis. Concluding remarks follow in Section 5.

2 Econometric framework

2.1 Potential Outcomes

My generalized quantile local projections (GQLP) framework builds upon the literature on local projections (Jordá 2005), quantile treatment effects in the presence of covariates (Powell 2020), and potential outcomes for time series (Rambachan and Shephard 2025).

Let $Y_{i,t}$ denote a scalar outcome variable and $Y_{j,t}$ the scalar treatment variable, both of which are part of a multivariate process $Y_t \in \mathbb{R}^k$. Let the vector X_t denote a finite history of the process, measurable with respect to the sigma-algebra generated by Y_t, i.e. X_t is $\mathcal{F}_t^Y$-measurable, where $\mathcal{F}_t^Y = \sigma\big(\{Y_s : s \leq t\}\big)$. Let the vector $W_t \in \mathbb{R}^k$ denote the assignments (or shocks in the terminology of macroeconomics). Throughout, capital letters represent random variables, while lowercase letters denote the fixed values these variables may take.

Following Rambachan and Shephard (2025) I define a potential outcomes process as a multivariate time series process that satisfies assumptions 1 and 2. The second adapts their sequentially probabilistic assignment condition: it conditions on the history of the outcomes

only, because the assignments are not observed, and it is stated for continuous assignments.

Assumption 1 (Non-anticipating Potential Outcomes)**.** *For each $t \geq 1$ and all sequences $\{w_s\}_{s\geq 1}$ and $\{w'_s\}_{s\geq 1}$ in $\mathcal{W}$,*

$$Y_t(w_{1:t}, \{w_s\}_{s\geq t+1}) = Y_t(w_{1:t}, \{w'_s\}_{s\geq t+1}) \text{ almost surely.}$$

I thus denote the time-t potential outcome as $Y_t(w_{1:t})$, where $Y_t(w_{1:t}) \in \mathcal{Y} \subseteq \mathbb{R}^k$.

Assumption 2 (Sequentially Probabilistic Assignment Process)**.** *For each t, conditional on $\mathcal{F}_{t-1}^Y$, W_t has a strictly positive conditional density $p_t(w \mid \mathcal{F}_{t-1}^Y)$ on $\mathcal{W}$. For discrete W_t, this corresponds to the overlap condition $0 < \Pr(W_t = w \mid \mathcal{F}_{t-1}^Y) < 1$.*

I focus on the dynamic effects on outcome variable i to a time t change in treatment variable j, thus I use the shortcut notation:

$$Y_{i,t+h}(w_j) := Y_{i,t+h}(W_{1:t-1}, W_{1:j-1,t}, w_j, W_{j+1:k,t}, W_{t+1:t+h}).$$

$Y_{i,t+h}(w_j)$ represents the potential outcomes that $Y_{i,t+h}$ would take had assignment $W_{j,t}$ been experimentally (exogenously) fixed to $W_{j,t} = w_j$.[2] The potential outcome is a random variable which depends on assignments up to time $t + h$. The set of potential outcomes includes the observed outcome, which in this shorter notation is denoted $Y_{i,t+h} \equiv Y_{i,t+h}(W_{j,t})$. Similarly, a potential outcome can be defined in terms of experimentally fixing $Y_{j,t} = y_j$ as $Y_{i,t+h}(y_j)$, in which case the observed outcome is $Y_{i,t+h} \equiv Y_{i,t+h}(Y_{j,t})$.

Before moving on to identification of quantile impulse responses, it is instructive to consider what assumptions are needed when the assignments W_t are unobservable. This is a standard setting in macroeconomics, where researchers often observe only a vector of endogenous variables Y_t but not the underlying assignments W_t. The typical model used in

[2]Throughout I focus on the potential outcomes for a fixed value of a variable in time t. As such the notation does not make explicit the fact that both the timing and the horizon of the treatment matter, for instance $Y_{i,t+h}(w_j)$ and $Y_{i,(t+1)+(h-1)}(w_j)$ may differ even though both occur in period $t + h$, because $Y_{i,t+h}(w_j)$ does not constrain the assignment in period $t + 1$ to equal w_j.

such settings is the structural vector autoregression. SVAR(1) assumes that the potential outcome process satisfies $A_0 Y_t(w_{1:t}) = w_t + A_1 Y_{t-1}(w_{1:t-1})$. The following assumptions are closely related but not identical to those used for identification in the SVAR literature, and although they can be applied in the SVAR setting they are more general and can be used to study identification in other (also nonlinear) settings.

Assumption 3 (Independent Assignments)**.** *Assignments are mutually independent across time and across the components of W_t. That is, each $W_{j,t}$ is independent of the collection of all other assignments $\{W_{i,s} : (i,s) \neq (j,t)\}$.*

Assumption 4 (Deterministic Potential Outcomes)**.** *Potential outcomes $Y_t(w_{1:t})$ are deterministic functions of the assignment sequence for all $t \geq 1$ and $w_{1:t} \in \mathcal{W}^t$.*

Assumption 5 (Partial Invertibility)**.** *The outcome $Y_{j,t}$ is a function only of the structural shock $W_{j,t}$ and the time-t observables X_t, i.e., $Y_{j,t} = g_j(W_{j,t}, X_t)$. Furthermore, the inverse $W_{j,t} = g_j^{-1}(Y_{j,t}, X_t)$ exists, and the range of $g_j(\cdot, x)$ is the same for every value x of the controls, so that every value of the treatment can be attained in every state. The controls are predetermined with respect to the shock: X_t is a vector of observed variables that is a function of the past assignments $W_{1:t-1}$ and of the time-t assignments other than $W_{j,t}$, but not of $W_{j,t}$ itself.*

Assumption 3 ensures that the assignments can be interpreted as structural shocks. It is a stronger version of the uncorrelatedness of structural shocks assumptions commonly used in the SVAR literature (if the structural shock process is jointly Gaussian the two assumptions are equivalent). Assumption 4 ensures that the potential outcomes (including the observed outcome) are a function of past and present structural shocks only, i.e. there are no "external" sources of randomness driving the potential outcomes.[3] Assumption 5 ensures that although the structural shock of interest $W_{j,t}$ is not observed, it can be recovered from observable data.

[3]The one exception is the initial condition: the system starts from an initial state drawn from its stationary distribution, independently of the assignments $\{W_t\}_{t\geq 1}$. This state summarises the assignments before the sample and is treated throughout as part of the past assignments.

This is a strong assumption, but it is weaker than the full invertibility assumption used in the SVAR literature – which requires that all the structural shocks W_t can be recovered. The predeterminedness part of assumption 5 requires that no control responds to the shock of interest within the period; with timing restrictions, it is satisfied by ordering the treatment after the controls (section 4). Together, these assumptions make the treatment as good as randomly assigned given the controls.

Lemma 1 (Conditional independence)**.** *Under assumptions 1, 3, 4 and 5, for every y_j and $h \geq 0$, $Y_{i,t+h}(y_j) \mid Y_{j,t}, X_t \sim Y_{i,t+h}(y_j) \mid X_t$.*

The proof is in appendix A1.1. In other words, since the treatment variable is assumed to be only a function of predetermined controls and an independent assignment, by conditioning on the appropriate variables the treatment is conditionally independent of the potential outcomes. Under these assumptions the structural mean impulse response is identified even though $W_{j,t}$ is not observed (proposition 2 in appendix A1.2).

2.2 Quantile Impulse Response definition

In what follows, I will assume that for a fixed treatment y_j and for each horizon h, $Y_{i,t+h}(y_j)$ has a structural quantile function (SQF) denoted $q_h(\tau \mid y_j)$. Notably, covariates X_t do not enter into this SQF, which distinguishes it from the conditional covariates SQF denoted $q_h(\tau \mid y_j, x)$, the τ quantile of $Y_{i,t+h}(y_j)$ given $X_t = x$.

Assumption 6 (Structural Quantile Function)**.** *For each h and y_j, the distribution of $Y_{i,t+h}(y_j)$ does not depend on t and there exists a structural quantile function $q_h(\tau \mid y_j)$ that is strictly increasing and left-continuous in $\tau \in (0, 1)$.*

Assumption 6 requires that the structural quantile function exists and does not change with time, ruling out models that feature structural breaks or other violations of stationarity. Each

potential outcome can be related to its structural quantile function as follows:

$$Y_{i,t+h}(y_j) = q_h(U_{i,t+h}(y_j) \mid y_j), U_{i,t+h}(y_j) \sim \mathsf{Uniform}[0,1].$$

$U_{i,t+h}(y_j)$ is responsible for heterogeneity of outcomes among time periods with the same treatment y_j. I refer to it as a rank variable as it determines the placement in the h-periods-ahead outcome distribution for a given treatment y_j. $U_{i,t+h}(y_j)$ contains information up to time $t+h$.

The goal of this paper is to identify the structural quantile impulse response, defined as:

$$\mathsf{QIR}_\tau(h) = \frac{\partial q_h(\tau \mid y_j)}{\partial y_j}. \tag{1}$$

If the SQF is linear i.e. $q_h(\tau \mid y_j) = \alpha_h(\tau) + \beta_h(\tau) y_j$, then the $\mathsf{QIR}_\tau(h) = \beta_h(\tau)$ does not depend on y_j. I discuss whether linearity of the SQF can be justified later. Importantly, I contrast this definition with the structural conditional quantile impulse response defined as:

$$\mathsf{cQIR}_\tau(h) = \frac{\partial q_h(\tau \mid y_j, x)}{\partial y_j}. \tag{2}$$

The QIR and the cQIR may differ even if the treatment and control variables are independent ($Y_{j,t} \perp\!\!\!\perp X_t$). Furthermore, the same observation $Y_{i,t+h}$ might fall below $q_h(\tau \mid y_j, x)$ but above $q_h(\tau \mid y_j)$ or vice versa.

The structural mean impulse response can be defined as:

$$\mathsf{IR}(h) = \frac{\partial \mathbb{E}[Y_{i,t+h}(y_j)]}{\partial y_j}, \tag{3}$$

and the structural conditional mean impulse response can be defined as:

$$\mathsf{cIR}(h) = \frac{\partial \mathbb{E}[Y_{i,t+h}(y_j) \mid X_t = x]}{\partial y_j}. \tag{4}$$

The expectation and quantile operators have different mathematical properties, which is why quantile and mean impulse responses behave differently under conditioning. To illustrate this, consider a simple data generating process: $Y = X_1 \cdot W + X_2$, where X_1, X_2 and W are all independently and identically distributed uniform random variables on $[0, 1]$ with W representing an unobserved shock. The linearity of the expectations operator means that $\mathbb{E}[Y \mid x_1] = x_1 \cdot \mathbb{E}[W] + \mathbb{E}[X_2]$ and $\mathbb{E}[Y \mid x_1, x_2] = x_1 \cdot \mathbb{E}[W] + x_2$. After taking the partial derivative with respect to x_1, both expressions yield $\mathbb{E}[W]$ making the IR and cIR identical. This result generalizes to any model where the treatment variable and other covariates enter additively[4], the linearity of expectations ensures that the additive component drops out when computing partial derivatives, regardless of whether it is conditioned on or not. In contrast, in general $q_{A+B}(\tau) \neq q_A(\tau) + q_B(\tau)$ for random variables A and B; the two sides are equal at every τ when A and B are comonotonic (Koenker 2005). After conditioning on both X_1 and X_2 they become constants, so $q_Y(\tau \mid x_1, x_2) = q_{x_1 W + x_2}(\tau) = x_1 \cdot q_W(\tau) + x_2$ by the affine property of quantiles. Taking the partial derivative with respect to x_1 then yields $\frac{\partial q_Y(\tau|x_1,x_2)}{\partial x_1} = q_W(\tau)$. However, after conditioning on $X_1 = x_1$ only, the conditional quantile $q_Y(\tau \mid x_1)$ cannot be written in separable form. Thus, the derivative $\frac{\partial q_Y(\tau|x_1)}{\partial x_1} = \frac{\partial q_{x_1 \cdot W + X_2}(\tau)}{\partial x_1}$ depends on the distribution of the random variable $x_1 \cdot W + X_2$ and in general differs from the conditional case. In effect, the coefficients on X_1 from quantile regressions on X_1 only versus on X_1 and X_2 will in general differ, even though X_1 and X_2 are independent. Adding covariates that are uncorrelated with the treatment to conditional mean models never changes the population coefficient on the treatment variable as per the Frisch-Waugh-Lovell theorem, but this theorem does not apply to quantile regression.

Which of the two objects is of interest depends on the question at hand. The cQIR measures how a shock shifts the distribution of the outcome in a given state of the economy, i.e. for given values of the controls. As such it is relevant for real-time risk assessment, for example when a central bank or a financial stability committee monitors GaR given current

[4]In models where the covariates do not enter linearly, Generalized IR functions that are a function of the covariates are needed and a simple linear model is misspecified.

conditions. This, however, requires modeling how the cQIR varies with the state, since a QLP that is linear in the controls estimates a single summary of the cQIRs across states, the treatment coefficient of a weighted least squares approximation of the conditional quantile function (Angrist et al. 2006) (section 3.1). On the other hand, the QIR measures how a shock shifts the distribution of outcomes across all periods, for example how much it raises the frequency and severity of recessions, which macro-prudential policy aims to reduce. Moreover, controls are usually chosen for identification, and changing them changes the interpretation of the cQIR but not of the QIR. The QIR is also distinct from the unconditional quantile effect of Firpo et al. (2009), who regress the recentered influence function of an unconditional quantile on the treatment and the controls. Their estimand measures how the quantile changes after a small shift in the distribution of the treatment, while the QIR compares the distributions of the outcome when the shock takes one value rather than another.

Comparing equations 1 and 3, the QIR and mean impulse response share the same structure but target different aspects of the outcome distribution. While the mean impulse response captures how treatment affects the expected value, the QIR describes how treatment affects specific quantiles. Both the QIR and IR should be interpreted as responses to shocks that cause a one-off time-t unit change in the treatment variable, consistent with the local projections literature. This interpretation differs slightly from SVAR impulse responses, which measure responses to unit variance innovations, so comparisons with SVAR impulse responses require scaling by an appropriate constant of proportionality (Plagborg-Møller and Wolf 2021).

A word of caution is in order when dealing with cumulative quantile impulse responses. To calculate cumulative impact on growth in the level of the variable of interest (e.g. Industrial Production IP_t) using local projections, the outcome variable is usually transformed to $Y_{i,t+h} = \log(IP_{t+h}) - \log(IP_{t-1})$. This is also the transformation used in this paper. This transformation is innocuous in the case of the mean impulse response as linearity of the expectations operator ensures that the cumulative effect equals the sum of period-by-period

effects. However, quantiles of sums generally do not equal sums of quantiles. For example, the effect on the median annual growth rate will not generally equal the sum of the effects on the 12 consecutive median monthly growth rates. Therefore, when $Y_{i,t+h}$ represents cumulative growth, the quantile impulse response describes how treatment affects the τ quantile of the h-periods-ahead cumulative growth distribution, not the sum of consecutive period effects. The identifying assumptions below, in particular rank similarity (assumption 7), are then assumptions about the cumulative outcome, and holding for a growth rate does not imply holding for its cumulative sum, or the reverse.

To motivate QIR analysis it is instructive to discuss in what class of models the QIR is not equal to the mean impulse response. In particular, I state a quantile invariance result that proves that in a rich class of models quantile impulse responses are the same for each conditional and unconditional quantile and equal to the mean impulse response. This result is useful for two reasons. First, it establishes a lower bound on model complexity necessary for a model to exhibit non-trivial quantile dynamics. This has implications for theorists interested in writing models which feature interesting quantile impulse responses consistent with empirical evidence. Second, it cautions against overly restrictive identifying assumptions for quantile impulse response identification, as that could lead to a paradox in which identification relies on assumptions that imply quantile invariance.

Suppose the process $Y_t \in \mathbb{R}^k$ admits a structural Wold representation of the form

$$Y_t = \sum_{\ell=0}^{\infty} \Psi_\ell W_{t-\ell},$$

where $\sum_{\ell=0}^{\infty} \|\Psi_\ell\| < \infty$ and $W_t \in \mathbb{R}^k$ are orthogonal innovations satisfying $\mathbb{E}[W_t] = 0$, $\mathbb{E}[W_t W_t'] = I_k$, and $\mathbb{E}[W_t W_s'] = 0$ for all $t \neq s$.[5] The marginal treatment effect of a shock in

[5] I use the term structural Wold representation for a linear moving average in the structural shocks; it coincides with the Wold representation of the process when the shocks are fundamental. The innovations dated before the start of the sample are those summarised by the initial state.

variable j at time t on variable i at horizon h is then

$$\frac{\partial \mathbb{E}[Y_{i,t+h}(w_j)]}{\partial w_j} = [\Psi_h]_{i,j}.$$

Additionally assume that the innovations are independent: the components of W_t are mutually independent and independent of W_s for all $s \neq t$. Gaussian innovations, $W_t \overset{iid}{\sim} \mathcal{N}(0, I_k)$, are a special case, because jointly Gaussian variables that are uncorrelated are independent. A stable linear SVAR with independent structural shocks, the leading case in macroeconomics, has such a representation.

Proposition 1 (Quantile Invariance)**.** *If $\{Y_t\}$ has a structural Wold representation with independent innovations and $[\Psi_0]_{j,j} \neq 0$, then for all quantiles $\tau \in (0,1)$, all horizons h and all controls X_t that satisfy assumption 5,*

$$\mathsf{cIR}(h) = \mathsf{IR}(h) = \mathsf{QIR}_\tau(h) = \mathsf{cQIR}_\tau(h)$$

As any purely nondeterministic, zero-mean covariance stationary process has a Wold representation, the most restrictive assumption in Proposition 1 is the independence of innovations. Critically, if the underlying data-generating process is non-linear, it may not admit a Wold representation with independent innovations even if the structural shocks driving the process are independent. For instance, processes with stochastic volatility have Wold representations but with innovations that are uncorrelated but not independent. As such, Proposition 1 tells us that breaking quantile invariance requires either nonlinearity or innovations that are uncorrelated but not independent. Since structural shocks are usually thought of as independent innovations, quantile impulse responses that differ across quantiles in practice require a nonlinear data generating process. This has a direct implication for the quantile local projections used in the literature, which are linear in the treatment and

the controls: if the data generating process is linear, all QIRs coincide with the mean impulse response, and if it is nonlinear, a linear QLP is generally misspecified. The proof of Proposition 1 follows from the linearity of the Wold representation and the independence of innovations. The full proof is in Appendix A1.4.

2.3 Quantile Impulse Response identification

If the observed treatment $Y_{j,t}$ is randomly assigned i.e. $U_{i,t+h}(y_j) \mid Y_{j,t} \sim U_{i,t+h}(y_j) \sim \mathsf{Uniform}[0,1]$, then a quantile local projection model $Y_{i,t+h} = q_h(U_{i,t+h} \mid Y_{j,t})$ estimated using a standard quantile regression restriction $P(Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}) = \tau$ identifies the QIR as defined in equation 1. In non-experimental settings typical in macroeconomics, an endogeneity problem arises because the realized treatment $Y_{j,t}$ is not randomly assigned. I address the endogeneity problem with an identification by controls strategy. In particular, I relax the assumption that $U_{i,t+h}(y_j) \mid Y_{j,t} \sim U_{i,t+h}(y_j)$ and replace it with $U_{i,t+h}(y_j) \mid Y_{j,t}, X_t \sim U_{i,t+h}(y_j) \mid X_t$.[6] This follows from lemma 1, since the rank variable is a function of the potential outcome. In other words, the treatment is randomly assigned conditional on the (observable) controls. As such the object of causal analysis is the quantile impulse response to a structural shock to the treatment variable.

The quantile local projections model with controls $Y_{i,t+h} = q_h(U^*_{i,t+h} \mid Y_{j,t}, X_t)$ estimated using a restriction $P(Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}, X_t) \mid Y_{j,t}, X_t) = \tau$ deals with the endogeneity issue, but estimates a different structural function $q_h(\tau \mid y_j, x)$ instead of $q_h(\tau \mid y_j)$. As such it estimates the cQIR defined in equation 2 instead of the QIR defined in equation 1. The addition of controls into the equation changes the interpretation of the model. In principle, the QIR could be obtained from the conditional quantile functions, by integrating the conditional distribution of the outcome over the controls and inverting it at each horizon. In practice this is hard, because the conditional quantiles must then have the right functional

[6] Note that this allows for the rank variable to have different distributions for different values of the controls X_t. I.e. the controls can help predict whether the outcome will be below/above its conditional (on treatment) quantile.

form in the treatment and in the controls, at every quantile and horizon. GQLP avoids this, since it uses the controls only in a first-stage regression and does not need a model of how they shift each quantile of the outcome. Note also that the conditional on controls rank variable $U^*_{i,t+h}(y_j, x)$ is distinct from $U_{i,t+h}(y_j)$. In particular, $U_{i,t+h}(y_j) = \Lambda(X_t, U^*_{i,t+h}(y_j, x))$ for some function Λ that depends on the fixed treatment and the horizon, but not time: $\Lambda(x, u) = F_h(q_h(u \mid y_j, x) \mid y_j)$, where $F_h(\cdot \mid y_j)$ is the distribution function of $Y_{i,t+h}(y_j)$.[7]

Exploiting control variables for causal identification while still modeling the conditional on treatment only SQF $q_h(\tau \mid y_j)$ is possible thanks to the Powell (2020) generalized quantile regression estimator, which explicitly distinguishes between treatment and control variables. One of the contributions of this paper is to adapt this cross-sectional framework to the time-series setting. In what follows I only consider identification by controls; Powell (2020) also considers identification using instrumental variables, which I leave for future work. To identify quantile impulse responses one more assumption is needed:

Assumption 7 (Rank Similarity)**.** *For each horizon h and quantile τ considered and all y_j, y'_j:* $\mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid Y_{j,t}, X_t] = \mathbb{P}[Y_{i,t+h}(y'_j) \leq q_h(\tau \mid y'_j) \mid Y_{j,t}, X_t]$.

The rank similarity assumption 7, condition 1D of Powell (2020) and a version of the assumption introduced by Chernozhukov and Hansen (2005), posits that, conditional on current observables, the rank of the potential outcome within its distribution does not systematically vary with different realizations of the treatment variable. In other words, if we know the current treatment and controls, whether a time period would have a high-rank or low-rank outcome does not depend on which treatment value we are considering. For example, if economic conditions suggest a period would experience an above-median outcome given one treatment value, those same conditions suggest it would also experience an above-median outcome for a different treatment value (though the median levels themselves would differ). It is a key assumption for identification, along with assumptions 3, 4 and 5, which taken together establish conditional (on controls) independence of the treatment assignment (lemma

[7] In Powell (2020) notation $U^*_d = \lambda_d(X, U_d)$, with starred rank variable being the unconditional one.

1).

In a time series, rank similarity means that the treatment can move the distribution of future outcomes but does not reorder the periods within it. Given what is known at time t, a period that would be in the lower tail of growth at one value of the treatment would also be there at another, because what puts it there, the state of the economy and the shocks that follow, acts in the same way whatever the treatment. Rank similarity does not rule out state-dependent effects. In the monotone transformation SVAR of section 3.1 an adverse shock does more damage in periods that are already weak, but because those periods are at the bottom of the distribution whatever the treatment, the ordering is preserved and rank similarity holds exactly (appendix A2.2.2). It fails when the treatment reorders periods, that is, when the effect of a shock depends on something that would not otherwise place a period in the lower tail. In the DSGE model of section 3.2, for example, a large adverse move can open a run equilibrium in fragile states, and the sunspot that then selects a run matters for the rank only when the move is large. There rank similarity holds only approximately, and section 3.2 shows that GQLP still recovers the QIR.

Before stating the moment conditions used to recover the QIR, I restate Theorem 1 of Powell (2020) in the time series setting. The proof of the theorem is in the appendix section A1.5.

Theorem 1. *Suppose Assumptions 1 through 7 hold. Then* $\forall h \in \{0, 1, 2, \ldots, H\}$ *and for each* $\tau \in (0, 1)$*:*

$$\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}, X_t] = \mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t],$$

$$\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})] = \tau.$$

The first equation in theorem 1 states that after conditioning on controls X_t, the treatment $Y_{j,t}$ does not provide additional information about the probability that the outcome is below its quantile function. The second equation in theorem 1 ensures that the quantile function is correctly scaled. Together, these equations imply that the conditional probability $\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t]$ is allowed to vary based on controls X_t, but in expectation it is

equal to the quantile level τ. When there are no control variables in the model (i.e. X_t is a constant), the two conditions in theorem 1 collapse into one standard quantile regression restriction $\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}] = \tau$. This restriction is used to estimate QIRs in the quantile local projections framework. As such, quantile local projections are a special case of generalized quantile local projections, corresponding to a setting where all the model variables are treatment variables and there are no controls. Therefore, GQLP "nests" the QLP framework.

Theorem 1 states that the structural quantile function satisfies these two restrictions. Identification also requires that no other function satisfies them. With the linear specification used below, this is the condition that the moment conditions have a unique root (assumption 9(e) in section 2.4), which appendix A2.2.3 checks in the monotone transformation SVAR of section 3.1, analytically when the SQF is linear and numerically otherwise.

2.4 Estimation

Theorem 1 provides the moment conditions needed for the estimation of the generalized quantile local projections. In particular, it implies two moment conditions for each horizon $h \in \{0, 1, 2, \ldots, H\}$ and quantile τ of interest:

$$\mathbb{E}\big[Y_{j,t}[\mathbf{1}\{Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})\} - \mathbb{P}(Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t)]\big] = 0,$$
$$\mathbb{E}\big[\mathbf{1}\{Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})\} - \tau\big] = 0,$$

where $\mathbf{1}\{\}$ is the indicator function that equals 1 if the condition in braces is true and 0 otherwise. Estimation is done separately for each horizon h and each τ as in the quantile local projections. For a given h and τ and assuming a linear specification $q_h(\tau \mid y_j) = \alpha_h(\tau) + \beta_h(\tau) y_j$, estimation proceeds in three steps:

1. Postulate a candidate $\tilde{\beta}_h(\tau)$. For each candidate $\tilde{\beta}_h(\tau)$ there exists an intercept $\tilde{\alpha}_h(\tau)$ such that $\mathbb{P}(Y_{i,t+h} \leq \tilde{\alpha}_h(\tau) + \tilde{\beta}_h(\tau) Y_{j,t}) = \tau$. This means that we need to search over the

slope coefficients only.

2. Given the pair $(\tilde{\alpha}_h(\tau), \tilde{\beta}_h(\tau))$, estimate a linear probability model (Logit or Probit could also be used, as in Powell (2020); the results below are for the linear probability model) for the event that $Y_{i,t+h} \leq \tilde{\alpha}_h(\tau) + \tilde{\beta}_h(\tau) Y_{j,t}$ as a function of a constant and the controls X_t. Save the predicted probabilities as $\hat{\tau}_{X_t}$.

3. $\hat{\beta}_h(\tau) = \operatorname{argmin}_{\tilde{\beta}_h(\tau)} \hat{m}' A \hat{m}$, where $\hat{m} = \frac{1}{T} \sum_{t=1}^{T} Y_{j,t} [\mathbf{1}\{Y_{i,t+h} \leq \tilde{\alpha}_h(\tau) + Y_{j,t} \tilde{\beta}_h(\tau)\} - \hat{\tau}_{X_t}]$.[8] With a single slope coefficient the model is exactly identified, so the weighting matrix A plays no role and $\hat{\beta}_h(\tau)$ makes $\hat{m}$ as close to zero as possible. Because $\hat{m}$ is a step function of $\tilde{\beta}_h(\tau)$, a local optimizer such as Nelder–Mead can stop on a flat step away from the minimum, so I minimize by grid search.[9]

As in Powell (2020), the binary outcome model may be misspecified; what identification requires is that its error, $\mathbf{1}\{Y_{i,t+h} \leq \alpha_h(\tau) + \beta_h(\tau) Y_{j,t}\} - \hat{\tau}_{X_t}$, be uncorrelated with the treatment $Y_{j,t}$ (in Powell's instrumental variable formulation the treatment plays the role of the instrument). With a linear probability model this holds if either the probability of the rank event or the conditional mean of the treatment is linear in the first-stage regressors (lemma 2 in appendix A1.6), and splines of the controls make both approximations more accurate. The same algorithm applies to a nonlinear specification $q_h(\tau \mid y_j) = \alpha_h(\tau) + f(y_j)' \beta_h(\tau)$, where f is a vector of known functions of the treatment, such as a polynomial or a spline: the vector $f(Y_{j,t})$ takes the place of $Y_{j,t}$ in the moment condition, and the search is over the vector $\beta_h(\tau)$.

The controls also reduce the variance of the GQLP estimates, but through a different channel than in QLP. In QLP part of the gain in precision comes from the change of the

[8] The sample moment vector is written as $\hat{m}$ (rather than the conventional g) to avoid a clash with the structural mappings g_j used elsewhere in the paper. Throughout, T denotes the number of periods for which the outcome $Y_{i,t+h}$, the treatment and the controls are all observed, which depends on h.

[9] I use a grid of step 0.05 within ± 3 of a starting value, followed by a grid of step 0.005 around the best point of the first grid. When several grid points attain the minimum, their midpoint is used. The starting value is the estimate of the default Nelder–Mead optimizer and, in the bootstrap, the full-sample estimate. In the empirical application the full-sample estimate is also compared with the minimizer over a wide grid, from -15 to 15 in steps of 0.05, and it is replaced by that minimizer, after the same refinement, when the criterion is lower there more than 0.1 away.

estimand, as the outcome is typically less dispersed given the controls. In GQLP the controls enter only through the first stage and do not change the estimand. Because the first stage is a least squares regression on the controls, the sample moment equals the covariance between the part of the treatment that the controls do not explain and the part of the rank event that they do not explain. Controls that help predict whether the outcome falls below its quantile shrink this second part, and with it the variance of the estimator, as controls that predict the outcome do in a linear regression. Controls that only predict the treatment shrink the first part and increase the variance.

This also guides the choice of controls. Their first role is identification: by assumption 5 they must contain everything, apart from the structural shock, that the treatment depends on. With timing restrictions these are the variables ordered before the treatment and enough lags to capture the predictable part of the treatment, but no variables that respond to the shock within the period. Beyond identification, additional controls can improve precision, but in short samples a large first stage overfits, and since rank similarity (assumption 7) is assumed conditional on the controls, each additional control can only make it more demanding. It is therefore advisable to choose minimal controls sufficient for satisfying the partial invertibility assumption 5.

Under stationarity and weak dependence, the GQLP estimator is consistent and asymptotically normal. Let $\theta = (\alpha_h(\tau), \beta_h(\tau))'$, let $\tilde{X}_t$ be the regressors of the linear probability model of step 2 (a constant and all the controls, possibly with transformations such as spline terms, so that $\tilde{X}_t$ carries the same information as X_t), let δ be its coefficients, and let $I_t(\theta) = \mathbf{1}\{Y_{i,t+h} \leq \alpha_h(\tau) + \beta_h(\tau) Y_{j,t}\}$. GQLP sets to zero the sample mean of

$$m_t(\theta, \delta) = \begin{pmatrix} Y_{j,t} \left[I_t(\theta) - \tilde{X}_t' \delta \right] \\ I_t(\theta) - \tau \\ \tilde{X}_t \left[I_t(\theta) - \tilde{X}_t' \delta \right] \end{pmatrix},$$

the two moment conditions of theorem 1, with the linear probability model in place of the

conditional probability, and the normal equations of that model. Let θ_0 be the true value and δ_0 the coefficients of the linear projection of $I_t(\theta_0)$ on $\tilde{X}_t$, let $G = \partial\,\mathbb{E}[m_t(\theta,\delta)]/\partial(\theta',\delta')$ at (θ_0,δ_0) be the Jacobian of the moments, let $v_t = Y_{j,t} - \tilde{X}_t'\pi$ be the residual of the linear projection of the treatment on $\tilde{X}_t$, and let φ_t be the density of $Y_{i,t+h}$ conditional on $(Y_{j,t}, X_t)$ at $\alpha_h(\tau) + \beta_h(\tau)Y_{j,t}$ and the true values.

Assumption 8 (Dependence)**.** *(a) (Y_t, X_t) is strictly stationary and β-mixing, with mixing coefficients β_ℓ of order $\ell^{-\nu}$ for some $\nu > p/(p-2)$ and $p > 2$. (b) $\mathbb{E}|Y_{j,t}|^{2p}$ and $\mathbb{E}\|\tilde{X}_t\|^{2p}$ are finite. (c) Conditional on $(Y_{j,t}, X_t)$, $Y_{i,t+h}$ has a density that is continuous and bounded by a constant that does not depend on $(Y_{j,t}, X_t)$, and the matrices $\mathbb{E}[\tilde{X}_t\tilde{X}_t']$ and $S = \mathbb{E}[\varphi_t\,(v_t, 1)'(1, Y_{j,t})]$ are nonsingular. (d) $\Omega = \sum_{\ell=-\infty}^{\infty} \mathsf{Cov}\big(m_t(\theta_0,\delta_0), m_{t+\ell}(\theta_0,\delta_0)\big)$ is finite.*

Assumption 9 (Regularity)**.** *(a) (θ_0,δ_0) is an interior point of a compact set Θ. (b) $q_h(\tau \mid y_j) = \alpha_h(\tau) + \beta_h(\tau)y_j$. (c) $Y_{i,t+h}$ is continuously distributed conditional on $(Y_{j,t}, X_t)$. (d) $\mathbb{E}\big\{Y_{j,t}\,[\mathbb{P}(I_t(\theta_0) = 1 \mid \tilde{X}_t) - \tilde{X}_t'\delta_0]\big\} = 0$. (e) (θ_0,δ_0) is the unique zero of $\mathbb{E}[m_t(\theta,\delta)]$ on Θ.*

Theorem 2 (Consistency and Asymptotic Normality)**.** *Suppose assumptions 1 through 9 hold for the given h and τ. Then $(\hat{\theta},\hat{\delta}) \to_p (\theta_0,\delta_0)$ and*

$$\sqrt{T}\begin{pmatrix}\hat{\theta}-\theta_0\\ \hat{\delta}-\delta_0\end{pmatrix} \to_d N\big(0,\ G^{-1}\Omega\,G^{-1\prime}\big),$$

which is Powell's $(G'WG)^{-1}G'W\Sigma WG(G'WG)^{-1}$ with Ω in place of Σ when the system is exactly identified.

Assumption 8 replaces the independent observations of Powell's theorem 4 with stationary, β-mixing data, and assumption 9 is his identification assumption for a linear probability model. Part (d) of assumption 9 is the condition, discussed above, that allows the first stage to be misspecified (lemma 2), and part (e) requires the moment conditions to have a unique root. Appendix A1.6 relates both assumptions to Powell's and proves the theorem, adapting his proof to dependent data. I calculate confidence intervals using moving block bootstrap

(Künsch 1989); the description of the algorithm is in the appendix section A1.7, together with a discussion of its validity.

3 Monte Carlo

The Monte Carlo study evaluates the estimators in two lab economies. The first, in section 3.1, is a monotone transformation SVAR featuring an ad hoc amplification mechanism, constructed so that assumptions 1 through 7 all hold exactly at every horizon. The second, in section 3.2, is the globally-solved nonlinear DSGE model of Gertler et al. (2019), in which the estimators' assumptions need not hold exactly, and which I use to check that GQLP recovers an asymmetric, nonlinear QIR from an endogenous observed treatment in samples of realistic length.

In both models I compare all estimators with the same benchmark: the slope that a linear quantile regression of the outcome on the treatment would find in a very long sample in which the treatment is assigned at random, independently of the state of the economy. I compute it with Algorithm 2 of appendix A2.1.2, and the structural quantile functions shown in the figures with Algorithm 1; appendix A2.1 describes both.

3.1 A Monotone Transformation SVAR

The first data generating process is a SVAR(1) in a latent index, observed through a nonlinear amplification map:

$$
\begin{aligned}
Z_{i,t} &= a_{1,11} Z_{i,t-1} + a_{1,12} Y_{j,t-1} + W_{i,t} \\
Y_{j,t} + a_{0,21} Z_{i,t} &= a_{1,21} Z_{i,t-1} + a_{1,22} Y_{j,t-1} + W_{j,t} \\
Y_{i,t} &= g(Z_{i,t}), \qquad g(z) = (1+\lambda)(z - z_0) - \frac{\lambda}{\kappa} \log\!\left(1 + e^{\kappa(z - z_0)}\right),
\end{aligned}
$$

where $W_{i,t}, W_{j,t} \overset{iid}{\sim} N(0,1)$ are unobserved independent structural shocks. The first two equations are an ordinary linear Gaussian SVAR(1) in the latent index $Z_{i,t}$ and the treatment $Y_{j,t}$, with $Z_{i,t}$ ordered first ($Z_{i,t}$ predetermined with respect to $Y_{j,t}$). The third equation is an observation equation: what the econometrician observes is not the index itself but a strictly increasing, concave transformation of it. If parameter $\lambda = 0$ the map g is affine and the dynamics of the observables are those of a standard linear SVAR. When $\lambda > 0$ the slope of the observation map,

$$g'(z) = 1 + \frac{\lambda}{1 + e^{\kappa(z - z_0)}} \in (1, 1 + \lambda), \tag{5}$$

is decreasing in z: it equals $1 + \lambda$ far below the threshold z_0 and 1 far above it. The same movement in the index therefore translates into a much larger movement in the observed outcome when the economy is already weak. This generates a left skewed ergodic distribution of $Y_{i,t}$ and QIRs that vary across quantiles. Throughout this subsection the outcome variable is the forwarded value $Y_{i,t+h}$.

parameter	$a_{1,11}$	$a_{1,22}$	$a_{1,12}$	$a_{1,21}$	$a_{0,21}$	λ	κ	z_0
value	0.8	-0.1	-0.25	-0.1	0.2	4	2	-1.5

Table 1: Model parameters used in the simulation.

Although this is not an economic model, to keep the discussion less abstract, think of $Y_{i,t}$ as output growth and $Y_{j,t}$ as the change in financial conditions (with positive values meaning tightening financial conditions). The latent index $Z_{i,t}$ is then the growth rate that the economy would deliver if credit intermediation always worked frictionlessly, and g summarizes the amplification that occurs once fundamentals deteriorate far enough for financial frictions to bind. The three parameters of the observation map have a direct reading. The amplification strength λ fixes how much harder a bad fundamental hits output than a good one helps it: at the calibration in Table 1 the marginal effect of the index on output is five times larger in a deep downturn than in a boom. The threshold z_0 locates the region in which the amplification is active; at $z_0 = -1.5$ it sits about 0.7 standard deviations of $Z_{i,t}$

below its unconditional mean, so the amplified region is visited in bad but not exceptional times. The smoothness κ governs how abruptly the economy passes from the unamplified to the amplified regime. The remaining parameters play the same role as in any SVAR: $a_{1,11}$ is the persistence of the fundamental, $a_{1,12}$ is the effect of tighter financial conditions on next period's fundamental, and $a_{0,21}$ makes financial conditions respond within the period to the state of the economy, which is what makes $Y_{j,t}$ endogenous and controls necessary. I use the amplification map as a simple reduced-form representation of the amplification mechanisms studied in the macrofinance literature (Kiyotaki and Moore 1997; Brunnermeier and Sannikov 2014; He and Krishnamurthy 2019; Gertler et al. 2019), including the DSGE model of section 3.2. In these models an adverse shock that hits when balance sheets are already weak is amplified through collateral constraints, fire sales and bank runs, while the same shock in good times has a much smaller effect.

Two properties make this example useful. First, assumptions 1 through 7 of section 2 hold exactly, at every horizon: the structural shocks are i.i.d. Gaussian and the potential outcomes are deterministic functions of the assignment sequence (assumptions 1 through 4), the treatment equation is additively invertible in $W_{j,t}$ (assumption 5), and, because the treatment shifts a single latent index that is passed through a strictly increasing map, the structural quantile function exists and the rank similarity assumption holds — in fact in its strongest, rank invariance, form (assumptions 6 and 7).[10] Second, the structural quantile function and the impulse responses are available in closed form (appendix A2.2.1), which I use in appendix A2.1.3 to check against the simulated benchmark.

In closed form, both the QIR and the cQIR equal the linear response of the latent index to the treatment, scaled by the slope of the amplification map at the point of the index distribution that the relevant quantile occupies. Conditioning on the controls removes the variance contributed by the current state of the economy, so the conditional lower quantile sits closer to the center of the index distribution, where the amplification is weaker. As a

[10]The formal rank similarity argument is in appendix A2.2.2.

result, when the economy is at its average state, the cQIR understates the lower-tail response, while at the median the two responses coincide. In sufficiently weak states the lower-tail cQIR exceeds the QIR in absolute value (equation (9)).

The treatment $Y_{j,t}$ is endogenous because of the contemporaneous term $a_{0,21}Z_{i,t}$ in its structural equation, $Y_{j,t} = g_j(W_{j,t}, X_t)$. That equation involves the latent index, which the econometrician does not observe, so the feasible control set consists of the observables $X_t = \{Y_{i,t}, Y_{i,t-1}, Y_{j,t-1}\}$. The functional form through which these observable controls enter is unknown to the econometrician, so I let them enter flexibly: each of $Y_{i,t}$ and $Y_{i,t-1}$ is expanded into a restricted cubic spline basis with five knots (four basis terms per variable), while the lagged treatment $Y_{j,t-1}$ enters linearly. In the GQLP estimator the spline terms enter the first stage of the Powell (2020) algorithm, where the binary event $\{Y_{i,t+h} \leq \tilde{\alpha}_h(\tau) + \tilde{\beta}_h(\tau)Y_{j,t}\}$ is modeled with a linear probability model in the controls. This combination works well in practice: the spline basis is flexible enough to approximate the unknown functional form of the rank probability in the controls, and the linear probability model is fast — unlike the logit or probit alternatives, which require iterative maximum likelihood estimation inside every evaluation of the GMM objective. As noted in section 2, this first stage need not be correctly specified: the treatment equation is linear in the latent index $Z_{i,t} = g^{-1}(Y_{i,t})$ and its lag, so the splines of $Y_{i,t}$ and $Y_{i,t-1}$ approximate the conditional mean of the treatment. In the QLP with controls the same spline terms enter as regressors.

I compare three feasible estimators, all with the forwarded outcome $Y_{i,t+h}$ as the dependent variable and all linear in the treatment: a QLP without controls; a QLP with the (spline) controls; and the GQLP with treatment $Y_{j,t}$ and the controls entering its first stage.

The markers in Figure 1 show the SQF at the first horizon, computed with Algorithm 1 of appendix A2.1.2, for three quantiles $\tau \in \{0.1, 0.5, 0.9\}$ over a grid of values of y_j. The three functions differ in level and in slope, and the vertical distance between them widens as the treatment tightens, which is the amplification mechanism at work: a tightening moves the lower quantile of growth much further than the median. Each function is concave, most

visibly at $\tau = 0.1$, because the lower quantile of the outcome distribution sits inside the amplification region for a wider range of treatments. How much this nonlinearity matters in practice is a quantitative question: over the range of treatments the model actually visits (two unconditional standard deviations either side of the mean), the slope of the $\tau = 0.1$ SQF varies between roughly -0.9 and -1.2, so the function is genuinely concave but mildly enough that a linear specification is an adequate description. Over that range the benchmark linear quantile regression is never more than 0.11 away from the SQF at $\tau = 0.1$, where the SQF falls by 5.0, and never more than 0.06 at the median, while at $\tau = 0.9$ the two are indistinguishable (appendix A2.1.3). I therefore work with linear specifications throughout, which keeps the comparison between the three feasible estimators about the estimand rather than about the functional form.

Outside of simulation settings, the structural shocks $W_{j,t}$ are usually unobserved, so researchers need to rely on the time-series of the observed endogenous model variables to estimate the SQF. Figure 1 compares the three feasible estimators of the first horizon SQF. The QLP without controls suffers from severe endogeneity bias: because financial conditions tighten when the fundamental deteriorates, the raw association between $Y_{j,t}$ and next period's growth overstates the causal effect by a factor of more than two and a half at $\tau = 0.1$, giving an average slope of -2.94 against a benchmark of -1.11. The QLP with controls removes that bias but targets the conditional SQF, which by equation (9) is a flatter object in the lower tail. Its average slope is -0.49. The GQLP targets the correct (unconditional on controls) SQF, while simultaneously addressing the endogeneity of the treatment in a controls-based identification strategy, and delivers an average slope of -1.03 at $\tau = 0.1$, close to the benchmark.

Figure 2 reports the estimated QIRs across horizons $h = 1, \ldots, 10$. The GQLP estimator, with the forwarded outcome $Y_{i,t+h}$ as the dependent variable, $Y_{j,t}$ as the treatment and the spline controls in its first stage, stays close to the benchmark. At the median and the upper quantile its mean bias never exceeds 0.01 (appendix A2.2 reports the mean bias and the root

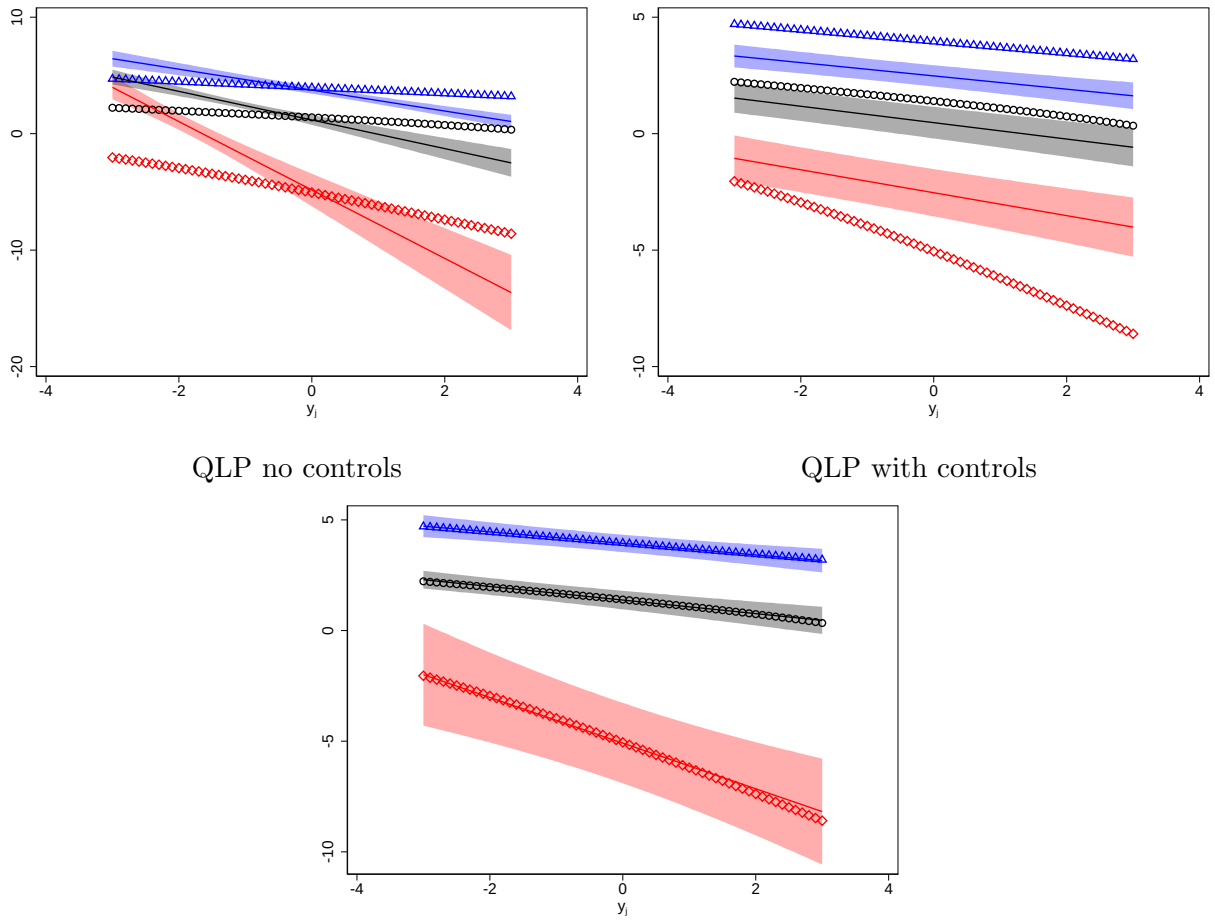
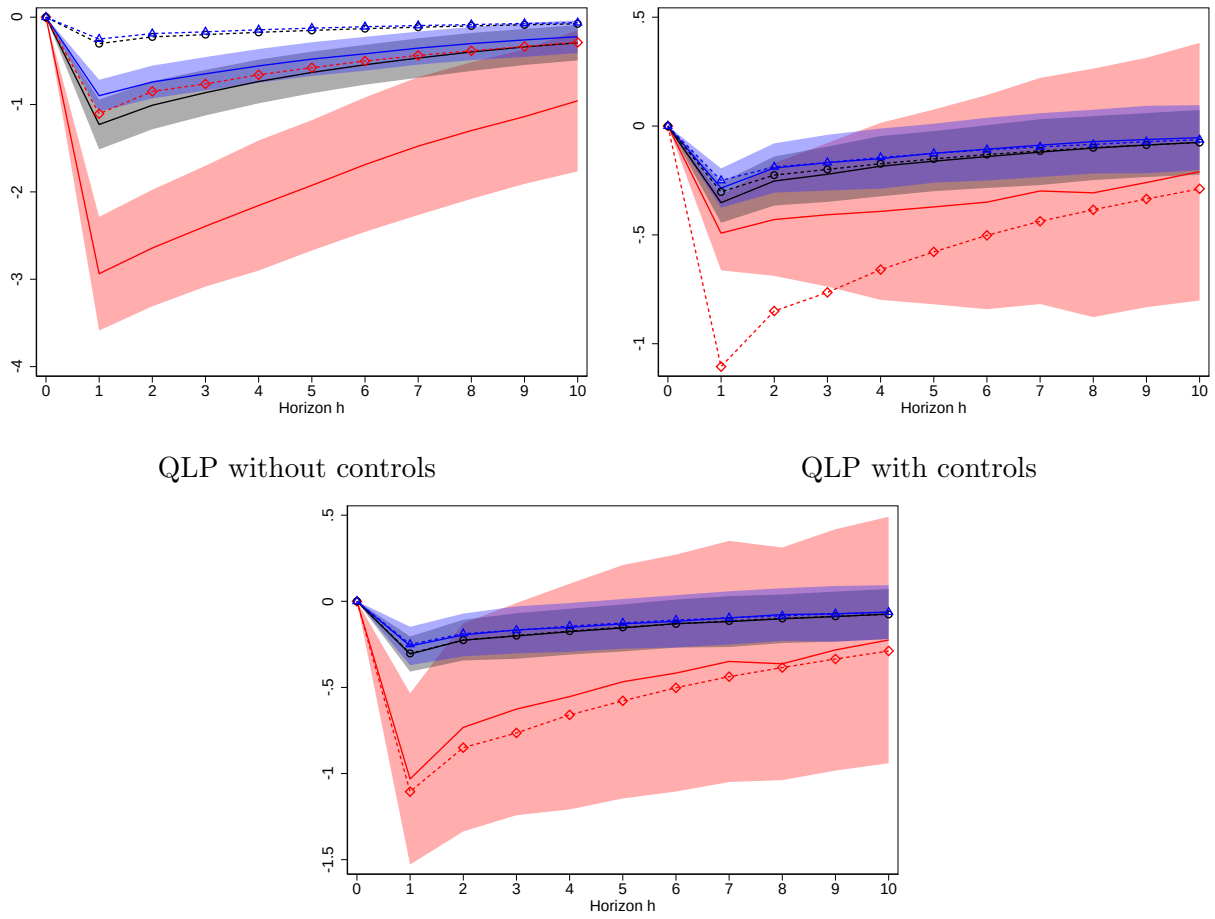


Figure 1: Monotone transformation SVAR: estimator comparison for the first horizon SQF $q_1(\tau \mid y_j)$. The diamonds, circles and triangles are the same across the three panels and show the quantiles of potential outcomes $Y_{i,t+1}(y_j)$ over a grid of values y_j, computed with Algorithm 1 of appendix A2.1.2, for quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. The three panels compare linear specifications of the three feasible estimators. QLP refers to the Quantile Local Projection estimated using the Koenker and Bassett (1978) estimator, GQLP is the Generalized Quantile Local Projection estimated using the Powell (2020) estimator. The regression coefficients used to plot the fitted SQFs are averaged estimates from $MC = 1{,}000$ simulation replications and time-series of length $T = 500$ (after dropping 1,000 burn-in observations). The QLP fits are evaluated at the sample means of the controls. Shaded areas are 68% Monte Carlo bands, the fitted value plus or minus one standard deviation of the fitted values across the replications.

mean squared error of every estimator). In the lower tail its average estimate is 7 percent below the benchmark in absolute value at the first horizon and 6–22 percent below it at longer horizons, where the response is small relative to the sampling noise. Most of this gap



QLP without controls

QLP with controls

GQLP

Figure 2: Monotone transformation SVAR: simulation results for the quantile impulse response of the forwarded outcome $Y_{i,t+h}$. The diamonds $\diamond$, circles $\circ$ and triangles $\triangle$ are the same across the three panels and show the benchmark, the population slope of a linear quantile regression of the potential outcome $Y_{i,t+h}(\tilde{y}_j)$ on a randomly assigned treatment $\tilde{y}_j$, computed with Algorithm 2 of appendix A2.1.2, for quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Solid lines show the results from the three feasible estimators. Results are averaged over $MC = 1{,}000$ simulation replications, with a time-series of length $T = 500$ (after dropping 1,000 burn-in observations). Y-axis plots $\hat{\beta}_h(\tau)$ and x-axis shows the horizon h. Shaded areas show the 68% Monte Carlo band of the estimator, equal to the estimate $\pm$ one standard deviation across the $MC = 1{,}000$ Monte Carlo iterations.

is the small-sample attenuation of lower-tail quantile regressions in samples of this length, not a property of GQLP: the infeasible quantile regression on the structural shock, which needs no controls, falls short of the benchmark by up to 15 percent, and the GQLP average

is within 0.05 of it at every horizon.[11] The QLP without controls is far off at every quantile and horizon. The QLP with controls understates the lower tail response by 0.61 at the first horizon, with the gap shrinking as the horizon grows. At the first horizon it also overstates the median and upper-decile responses, by 16 and 14 percent. The shaded areas display the Monte Carlo dispersion of the estimators: GQLP pays for targeting the unconditional object with a higher variance in the lower tail (a standard deviation across replications of 0.50 at the first horizon, against 0.17 for the conditional projection and 0.70 for the infeasible regression on the shock). At the first horizon, where the bias of the conditional projection is largest, GQLP has the lower root mean squared error (0.50 against 0.64); from the second horizon on, the conditional projection has the lower root mean squared error, but it estimates a different object. At the median and the upper quantile the two estimators are similarly precise.

		Nominal Level		
Quantile	Horizon	68 %	90 %	95 %
0.1	1	73%	93%	96%
	5	68%	91%	95%
	10	72%	90%	96%
0.5	1	71%	90%	95%
	5	73%	92%	96%
	10	73%	92%	97%
0.9	1	71%	91%	95%
	5	70%	92%	96%
	10	73%	91%	97%
Average coverage		71.6%	91.5%	95.9%

Table 2: The table reports the coverage of moving block bootstrap confidence intervals for the GQLP estimator in the monotone transformation SVAR, for three quantiles, three horizons and three nominal confidence levels. The coverage was computed in a Monte Carlo simulation with $MC = 500$ repetitions, sample size $T = 500$ and $B = 1{,}000$ bootstrap repetitions, with the spline controls of Figure 2 in GQLP's first stage. The true QIR is the benchmark of Figure 2. The block size used in the bootstrap procedure was the same as in the empirical results section and equal to $M = 7$.

Table 2 reports the coverage of the moving block bootstrap confidence intervals (CIs) for three horizons $h \in \{1, 5, 10\}$ for the GQLP estimator in this data generating process. The

[11] Appendix A2.2.4 repeats the experiment with samples of $T = 10{,}000$. The shortfall of GQLP then disappears while the bias of QLP with controls does not shrink.

algorithm used to compute the CIs is the same as the one used in the empirical section and is described in the appendix section A1.7. Coverage was calculated by computing the CIs for the QIR at three selected horizons and three quantiles of interest $\tau \in \{0.1, 0.5, 0.9\}$ over 500 Monte Carlo simulation repetitions, and then recording the percentage of repetitions in which the true value lay inside the CI. The empirical coverage is close to the nominal levels at all quantiles and horizons and, where it deviates, errs on the conservative side: the average coverage across cells is 71.6%, 91.5% and 95.9% against nominal levels of 68%, 90% and 95%. The bootstrap standard deviation is on average about 7% larger than the Monte Carlo standard deviation of the estimator, consistent with the result of Hahn and Liao (2021) that intervals built from the bootstrap standard deviation are asymptotically conservative.

3.2 Nonlinear DSGE model

"A Macroeconomic Model with Financial Panics" of Gertler et al. (2019) is a fully micro-founded nonlinear DSGE model that features bank panics and financial accelerator mechanisms. It is solved globally and so it can be used as a nonlinear data generating process by sampling random innovations. In the model households and bankers interact through capital accumulation, adjustment costs, and an agency problem. Capital is produced subject to convex adjustment costs, and households can hold capital directly but incur a management cost. Bankers finance their operations with their net worth and household deposits, but face diversion risk and potential bankruptcy if their net worth turns zero or negative. The only exogenous uncertainty in the model is a capital-quality shock, while endogenous bank runs are triggered by sunspot disturbances whenever liquidation prices fall sufficiently—reflecting households' relative inefficiency at managing capital—to push bank net worth below zero. Although the steady-state is free of bank runs, a sequence of adverse capital-quality shocks can erode net worth and open the door to panic equilibria. Appendix A2.3.1 shows how a sequence of negative capital-quality shocks can move the economy from the steady state to a bank run. In the simulations the sunspot is drawn independently with probability $\Pi = 0.15$, the value

under which the model is solved, so that runs occur in about 1 percent of quarters.

I treat growth in capital quality, $Y_{j,t} = \xi_t - \xi_{t-1}$, as the observed treatment. In the model capital quality scales the value of the assets on bank balance sheets, so $Y_{j,t}$ is a shock to bank asset quality, a model counterpart of the financial risk shocks studied in section 4. Capital quality follows $\xi_t = 1 - \rho + \rho\,\xi_{t-1} + \varepsilon_t$ with $\rho = 0.7$, so the treatment equation is

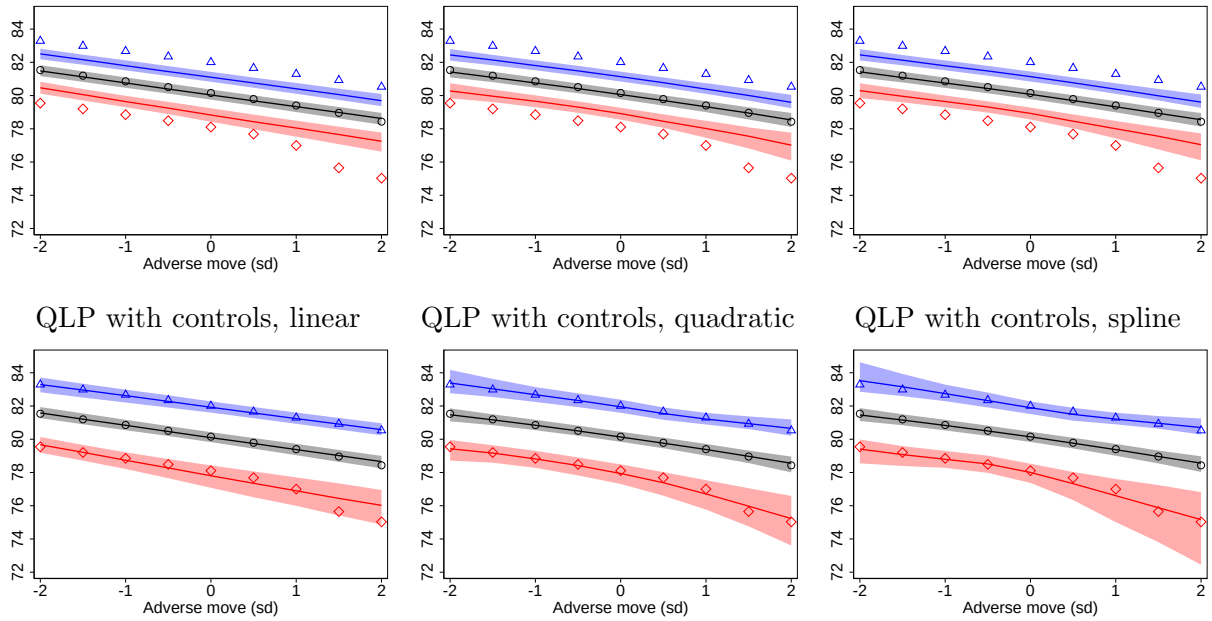
$$Y_{j,t} = (\rho - 1)(\xi_{t-1} - 1) + \varepsilon_t.$$

The treatment is invertible in the capital-quality innovation ε_t with a unit slope, so assumption 5 holds exactly whenever ξ_{t-1} is among the controls. It is also endogenous: its mean-reverting component is predictable from the lagged state, which also predicts future output. The sunspot does not enter the treatment, so a run at date t is a bad *outcome* at a given treatment. Counting the sunspot as a second structural assignment, assumptions 3 and 4 hold. Rank similarity (assumption 7) is not imposed and holds only approximately.[12] The outcome is log output, $\log Y_{i,t+h}$, for $h = 0, \ldots, 12$ quarters; appendix A2.4 repeats the experiment with cumulative growth, $\log Y_{i,t+h} - \log Y_{i,t-1}$, the transformation used in section 4. All responses are reported in percent per one-standard-deviation adverse (negative) move in $Y_{j,t}$.

The markers in Figure 3 show the structural quantile function after one quarter, computed with Algorithm 1 of appendix A2.1.2. The median and the upper decile are close to linear. The lower quantiles bend sharply once the adverse move exceeds about one standard deviation: such moves open a run equilibrium in fragile states, and with probability Π the sunspot selects it. The QIR is correspondingly asymmetric: over the first six quarters, the benchmark response of the $\tau = 0.1$ quantile (Figure 4) is about 1.3 times the median response.

I simulate 500 samples of $T = 320$ quarters, matching the length of the postwar quarterly sample. I compare three feasible estimators: 1) QLP without controls; 2) QLP with the controls $X_t = \{\xi_{t-1}, N_{t-1}, K_{t-1}, \log Y_{i,t-1}\}$, comprising lagged capital quality, bank net worth,

[12] The Spearman rank correlation between potential outcomes at treatment values two standard deviations above and below the mean is between 0.88 and 0.98 across horizons.



QLP with controls, linear | QLP with controls, quadratic | QLP with controls, spline

GQLP, linear | GQLP, quadratic | GQLP, spline

Figure 3: GKP model: estimates of the structural quantile function of $100 \times \log Y_{i,t+1}$ with linear (left), quadratic (middle) and spline (right) specifications in the treatment, for QLP with controls (top row) and GQLP (bottom row). Markers show the structural quantile function for $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$, computed with Algorithm 1 of appendix A2.1.2, and the horizontal axis is the adverse move in capital-quality growth, in standard deviations. Solid lines are medians of the fitted values across $MC = 500$ replications of length $T = 320$, and shaded areas, the 68% Monte Carlo bands, run from the 0.16 to the 0.84 quantile across replications. The QLP fits are evaluated at the sample means of the controls. Appendix A2.3.3 shows the shapes of the fits, each shifted to the structural quantile function at zero (Figure 12).

capital, and output, entered linearly; and 3) GQLP, which models the rank event in the first stage using a linear probability model with a restricted cubic spline (five knots) in ξ_{t-1}, the control required by the treatment equation.

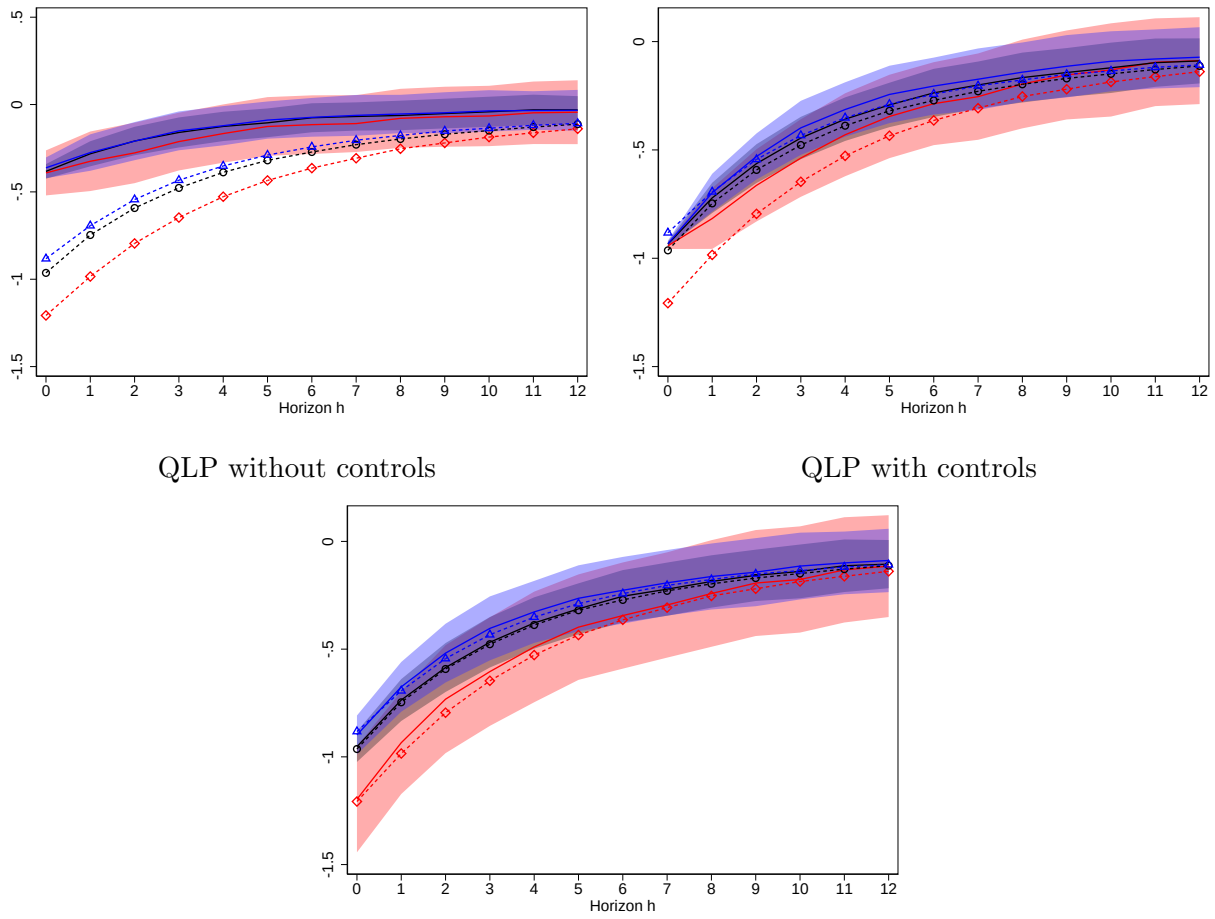
Because the lower quantiles of the structural quantile function bend, I first estimate the function itself, with a linear, a quadratic and a spline specification in the treatment. Figure 3 compares the fitted functions after one quarter. Linear GQLP misses the bend in the lower tail: at a two-standard-deviation adverse move its fitted $\tau = 0.1$ quantile is one point too high. With a quadratic or a spline GQLP follows the bend: the medians of its fits are within 0.4 points of the structural quantile function at every point of the grid, and between no move

and a two-standard-deviation adverse move its fitted $\tau = 0.1$ quantile falls by 2.7 points with the quadratic and 2.8 with the spline, against 3.1. QLP with controls misses the $\tau = 0.1$ and $\tau = 0.9$ quantiles with all three specifications. It estimates quantiles of output given the controls, and output is less dispersed given the controls than across states: with no move, its $\tau = 0.1$ and $\tau = 0.9$ quantiles are 2.2 points apart, against 3.9. Its $\tau = 0.1$ quantile also bends too little, falling by only 1.9 points with the quadratic or the spline, and at a two-standard-deviation adverse move it is 2 points too high.

To simplify the exposition, I assume a specification linear in the treatment variable for recovering the QIR. Figure 4 reports the results. QLP without controls recovers only about a third of the response at every quantile and horizon. On impact at $\tau = 0.1$ it gives -0.39 percent against a benchmark of -1.21: the predictable component of capital-quality growth moves with the state of the economy, and without controls it contaminates the estimates.

GQLP stays close to the benchmark at every horizon. At $\tau \in \{0.1, 0.5, 0.9\}$ its mean bias never exceeds 0.063 in absolute value, less than that of the infeasible quantile regression on the structural shock (0.087), and its Monte Carlo standard deviation is between 0.8 and 1.1 times that of the infeasible regression (appendix Table 11). It also reproduces the asymmetry: on impact the $\tau = 0.1$ response is 1.26 times the median response, against 1.25 in the benchmark.

QLP with controls is accurate at the median but misses the asymmetry on impact. Its lower-decile response (-0.94) is essentially equal its median response (-0.93), whereas the benchmark ratio is 1.25 and the GQLP ratio 1.26. On impact, output is close to a deterministic function of the controls and the treatment, except in the rare quarters in which a run occurs. A quantile regression that is linear in the controls fits the no-run response and treats the run quarters as outliers. GQLP instead models only the quantile of the outcome given the treatment. QLP with controls is, however, more precise than GQLP, especially on impact: as in the SVAR, targeting the unconditional quantile costs some efficiency in the lower tail. Appendix A2.3 reports the responses over the whole range of quantiles.



QLP without controls

QLP with controls

GQLP

Figure 4: GKP model: simulation results for the QIR of log output to a one-standard-deviation adverse move in capital-quality growth, in percent. The diamonds, circles and triangles show the benchmark, the population slope of a linear quantile regression of the potential outcome on a randomly assigned treatment, computed with Algorithm 2 of appendix A2.1.2, for $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Solid lines are the three feasible estimators, averaged over $MC = 500$ replications of length $T = 320$ (after 600 burn-in quarters). Shaded areas are the 68% Monte Carlo band, the estimate $\pm$ one standard deviation across replications.

4 Empirical Results

I apply the GQLP methodology to reexamine how financial risk shocks affect the distribution of industrial production growth in the United States. The empirical exercise follows an established literature that documents how adverse financial shocks disproportionately increase downside risks to growth (Adrian et al. 2019; Chavleishvili and Manganelli 2024; Loria et al.

2025). My choice of financial risk variables – capturing credit and volatility risk – is motivated by both theoretical considerations and empirical precedents. Theoretically, tightening financial conditions can amplify economic downturns through multiple channels including credit market frictions that constrain firm investment when external finance premiums rise (Gilchrist and Zakrajšek 2012), and uncertainty-driven real options effects that delay irreversible investments (Bloom 2009). Previous studies have used similar financial indicators to identify these channels, with Gilchrist and Zakrajšek (2012) focusing on the excess bond premium and Bloom (2009) focusing on stock market volatility. By contrasting results from GQLP and QLP using these indicators, I demonstrate that distinguishing between conditional and unconditional quantiles is both statistically and economically important for understanding the causal drivers of growth-at-risk.

My monthly dataset covers the US economy during the period between April 1973 and August 2026 for credit risk ($T = 641$) and between April 1986 and August 2026 for volatility risk ($T = 485$), the longest samples the data allow.[13] All the data used is publicly available, with the majority of it contained in the FRED-MD database published by the St. Louis Fed (McCracken and Ng 2015). I use monthly data for a larger sample size, with Industrial Production as the dependent variable[14]. In particular, the dependent variable $Y_{i,t+h}$ is defined as the h-months cumulative log growth rate $Y_{i,t+h} = 100\left[\log(IP_{t+h}) - \log(IP_{t-1})\right]$. I multiply the log growth rates by 100 to interpret the QIR in terms of percentage points. I normalize the treatment variable $Y_{j,t}$ to interpret the QIRs as responses to a one standard deviation change.

The first treatment variable $Y_{j,t}$ I consider measures movements in credit risk. I will refer to this variable as credit risk and I define it as the first-difference of the monthly Excess Bond Premium (EBP) of Gilchrist and Zakrajšek (2012), i.e. $Y_{j,t} = EBP_t - EBP_{t-1}$. The EBP is the residual of corporate bond credit spreads that cannot be explained by movements

[13]The consumer price index for October 2025 was not published because of the federal government shutdown, so the months from October 2025 to January 2026, whose controls include the inflation rate of October or November 2025, drop out of the estimation.

[14]Appendix section A3.3.7 contains results for the Unemployment rate as the dependent variable.

in expected default risk, as such it measures the investor sentiment or risk appetite in the corporate bond market.

The second treatment variable $Y_{j,t}$ I consider measures the volatility risk premium in the equity markets, defined as the difference between realized and implied volatility of the S&P500 index. I will refer to it as volatility risk for short. I compute realized volatility by computing the annualized standard deviation of daily log returns (based on close prices) in each month. I use the VIX as a measure of implied volatility[15]. The VIX is the option market's forecast of volatility over the next 30 days, so I compare realized volatility in month t with the average VIX in month $t-1$, both in annualized percentage points. If option markets are efficient, implied volatility should be an efficient forecast of future volatility; it should subsume the information contained in all other variables in the market information set in explaining future volatility. Thus, $Y_{j,t} = realized_t - implied_{t-1}$ captures realized volatility that was unexpected by the financial markets (Christensen and Prabhala 1998).[16]

I order the financial risk variable after macroeconomic variables but before financial markets and monetary policy variables. This assumes that financial conditions are affected contemporaneously by macroeconomic shocks but respond with a lag to shocks to monetary policy. Assuming that financial variables adjust quicker than real variables is justified by the speed at which financial markets respond to news and is a common assumption in the macroeconomic literature (Sims 1980; Christiano et al. 1996; Bloom 2009; Gilchrist and Zakrajšek 2012; Chavleishvili and Manganelli 2024). My variables are ordered as follows: {consumption growth, investment growth, industrial production growth, inflation, financial risk variable $Y_{j,t}$, S&P500 monthly return, change in the ten-year (nominal) Treasury yield, change in the effective (nominal) federal funds rate}. This ordering implies that the controls vector X_t must include the contemporaneous values of the four variables ordered before

[15]The VIX series of FRED-MD is option-implied from January 1986; before that it is the synthetic series of Bloom (2009), which is built from the realized volatility of the S&P500. The volatility risk sample therefore starts in 1986.

[16]Note that this sign convention is the negative of the volatility risk premium as it is usually defined (implied minus realized volatility); with the convention used here, positive values of $Y_{j,t}$ correspond to adverse volatility surprises, i.e. months in which realized volatility exceeded what option markets had priced in.

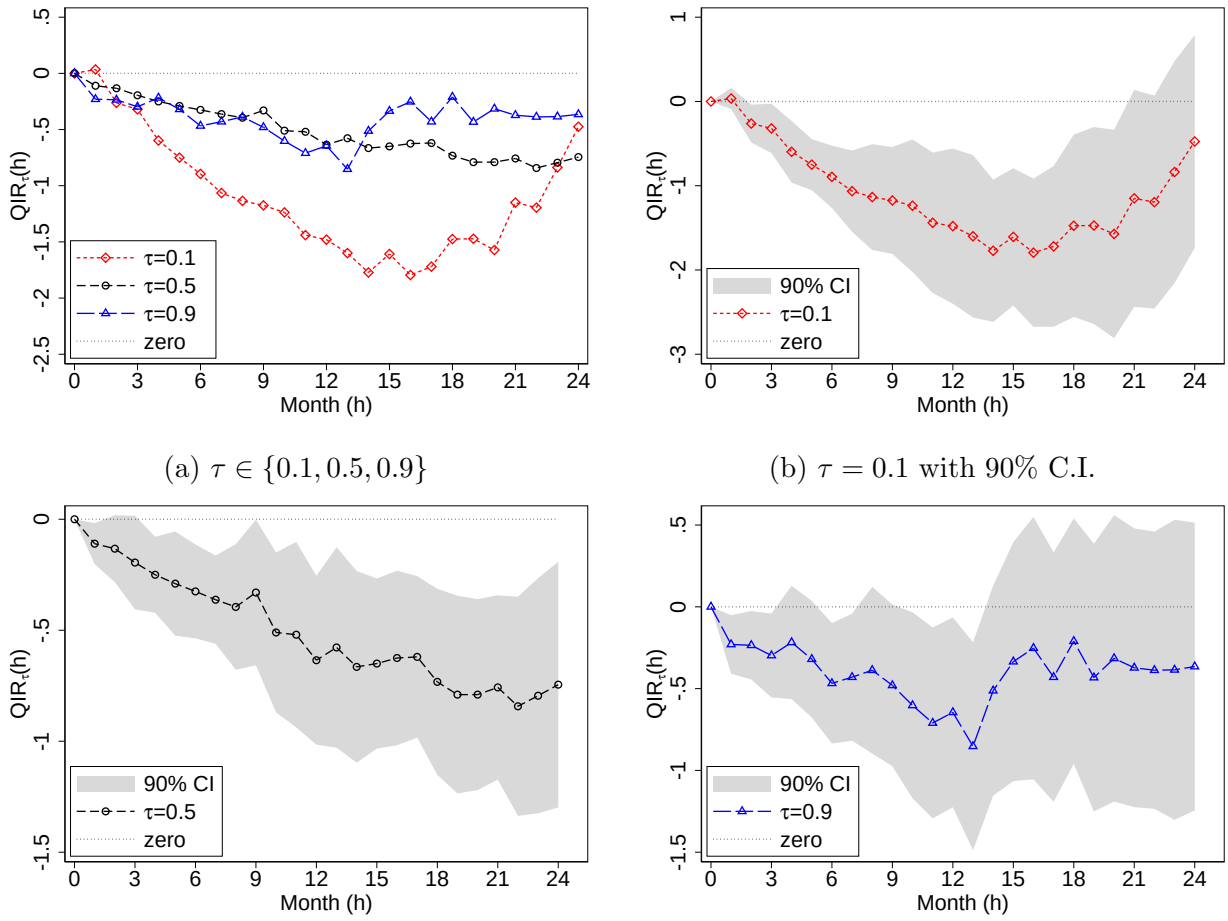
the treatment variable $Y_{j,t}$, and that these variables do not respond to the financial risk shock within the month, which is the predeterminedness part of assumption 5. Additionally, to control for the broad state of the economy in the recent past, I include the first two lags of all eight variables contained in my ordering in X_t. Both treatments are stationary transformations of persistent variables, whose predictable part depends on the lagged level, so I also include the lagged level of the financial risk variable: the excess bond premium for credit risk, and realized volatility and the VIX for volatility risk. In short, my timing restriction assumption allows for financial conditions to adjust within the period to consumption growth, investment growth, industrial production growth and inflation, but not to the stock market return, changes of the Treasury yields and changes to the federal funds rate. For volatility risk the ordering relative to the stock market return deserves a comment, because realized volatility and the return in month t are computed from the same daily returns. Ordering volatility first is consistent with the volatility feedback effect, in which news of higher volatility lowers stock prices within the period (French et al. 1987; Campbell and Hentschel 1992). To test the robustness of the conclusions to the timing restriction chosen, I report results with the treatment variable ordered first and last in the appendix section A3.3.2.

Throughout, I focus on three quantiles $\tau \in \{0.1, 0.5, 0.9\}$. The $\tau = 0.1$ quantile is of primary interest as it measures downside-risk. I also report results for a richer set of quantiles for four selected horizons, including the $\tau = 0.05$ quantile corresponding to the usual definition of GaR. To simplify the analysis I assume a linear specification for the SQF $q_h(\tau \mid y_j) = \alpha_h(\tau) + \beta_h(\tau) y_j$; this ensures that the $\mathsf{QIR}_\tau(h) = \beta_h(\tau)$ does not depend on y_j.[17] I choose a 90% confidence level for reporting the moving block bootstrap confidence intervals, which are computed using a block length of 7 and 1,000 bootstrap replications.[18]

Figure 5 shows the recovered QIRs of industrial production to a shock which increases credit risk by one standard deviation. The upper-left panel in Figure 5 plots the QIRs for

[17] Appendix section A3.3.1 contains results with a non-linear in treatment SQF.

[18] Appendix A1.7 explains the choice of the block length, and appendix A3.3.4 reports bands with blocks of 4, 12 and 24 months.



(a) $\tau \in \{0.1, 0.5, 0.9\}$

(b) $\tau = 0.1$ with 90% C.I.

(c) $\tau = 0.5$ with 90% C.I.

(d) $\tau = 0.9$ with 90% C.I.

Figure 5: Cumulative response of Industrial Production (in % pts.) from a shock that increases credit risk by one standard deviation, plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the quantile impulse response. Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Note that the impact response (horizon $h = 0$) is by assumption zero, given my timing restrictions.

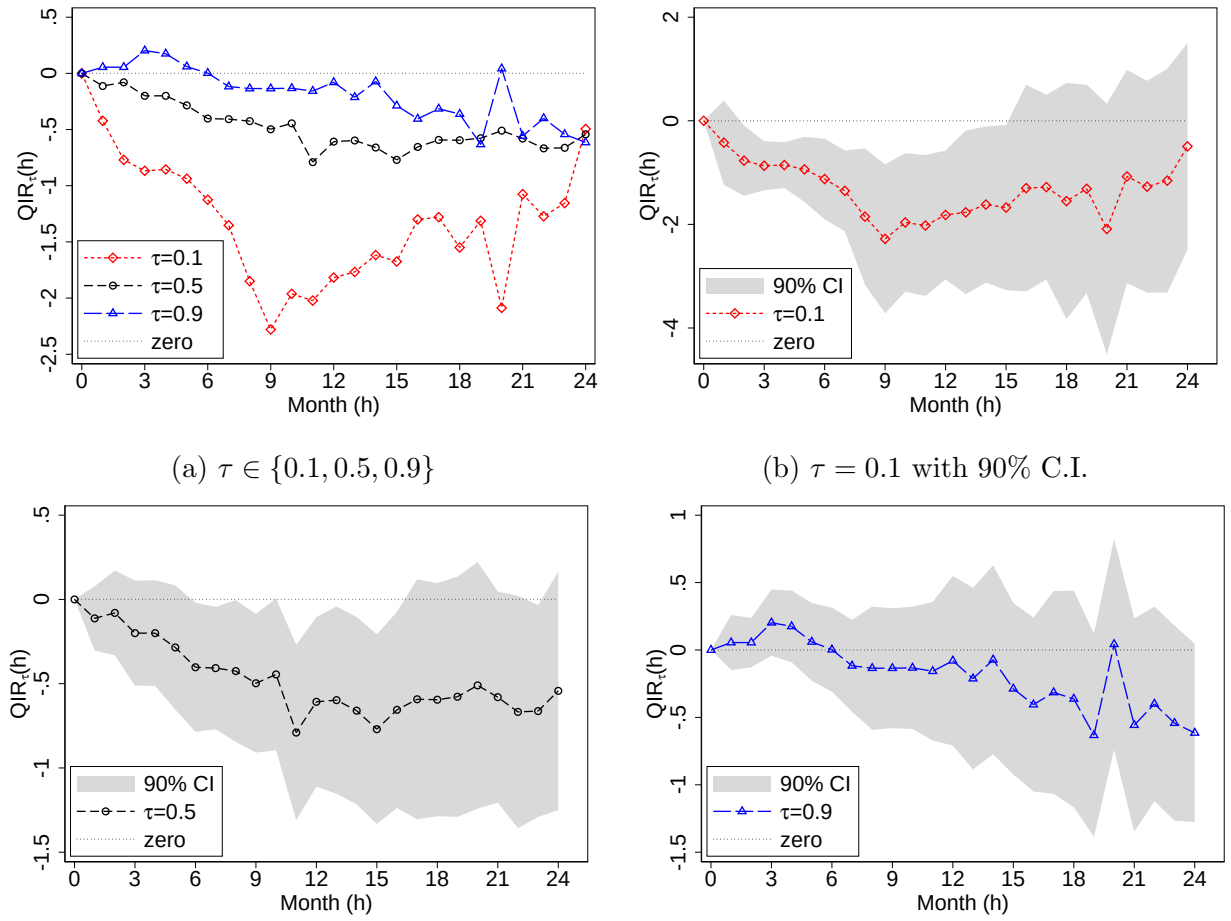
the three quantiles on the same axis. It is clear that the response at the $\tau = 0.1$ quantile is much more pronounced than the response at the other quantiles considered. This is a feature of the data and not of the model, as nothing is restricting the responses of lower quantiles to be lower than those of the upper quantiles. For instance, a shock that lowers the variance of a distribution would give rise to positive QIRs for quantiles below the median and negative QIRs for quantiles above the median.

My findings suggest economically large and statistically significant (at 90% confidence

level) declines in the $\tau = 0.1$ quantile of cumulative growth of up to 1.8 percentage points. The median losses ($\tau = 0.5$) are considerably smaller, at most 0.84 percentage points. The upside-risk response ($\tau = 0.9$) is similar to the median response over the first year, except that the effect is not statistically significant beyond the thirteen-month horizon, and it is smaller than the median response in the second year. The estimation uncertainty measured by the moving block bootstrap confidence intervals increases with the horizon, it is also higher for the $\tau = 0.1$ quantile than the median and the $\tau = 0.9$ quantile.

The four panels of Figure 15 in appendix A3.1 report results of the same model estimated for a richer set of quantiles (from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments) for four fixed horizons (6 months, 1 year, 18 months and 2 years). It shows that at all four of these horizons the slope of the structural quantile function is more negative for quantiles below the median than at the median, with one exception, the $\tau = 0.1$ quantile at the two-year horizon. The effect of financial shocks on growth is statistically significant but not for all quantiles. Quantiles below the median are more affected by financial shocks and the effect is more likely to be statistically significant even though it is estimated less precisely than the effect around the median growth scenario. The confidence intervals for the GaR $\tau = 0.05$ quantile are considerably wider than for the $\tau = 0.1$ quantile: with about 30 observations per horizon below the $\tau = 0.05$ quantile, the point estimates at this quantile are imprecise, which is why I focus on the $\tau = 0.1$ quantile as the measure of downside risk.

Figure 6 shows the results of estimating the same model but using volatility risk in place of credit risk as the treatment variable. Comparing Figure 6 to Figure 5 suggests that the relationship between volatility risk shocks and growth is similar to the relationship between credit risk shocks and growth. The median responses are similar in timing and magnitude. The lower tail responds faster to volatility risk, by 2.3 percentage points after nine months against 1.2 for credit risk, but from about the one-year horizon on the two lower-tail responses are close, although the band for volatility risk is wider and excludes zero only up to the fifteenth month. They differ in the upper tail: credit risk shocks lower the upper decile about



(a) $\tau \in \{0.1, 0.5, 0.9\}$

(b) $\tau = 0.1$ with 90% C.I.

(c) $\tau = 0.5$ with 90% C.I.

(d) $\tau = 0.9$ with 90% C.I.

Figure 6: Cumulative response of Industrial Production (in % pts.) from a shock that increases volatility risk by one standard deviation, plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the quantile impulse response. Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Note that the impact response (horizon $h = 0$) is by assumption zero, given my timing restrictions.

as much as the median, while volatility risk shocks leave it nearly unchanged during the first year, and its response to volatility risk is not statistically significant at any horizon. Both volatility and credit risk affect down-side more than upside-risk. The similarities are striking considering the fact that the sample correlation coefficient between these two variables is only 0.3. These findings suggest either the existence of a common non-linear propagation mechanism (as argued for by Loria et al. 2025) or the fact that it is the overall financial conditions – of which credit and volatility are both components – that have an asymmetric

effect on the distribution of output growth.

As before, I report the results for more quantiles at four fixed horizons for the volatility risk in Figure 16 (appendix A3.1). Figures 15 and 16 are similar below the median, except at $\tau = 0.05$, where the responses to volatility risk are larger at three of the four horizons and less precisely estimated. Again, this suggests that the relationship between financial shocks and the left-tail of growth depends little on whether the shocks pertain to credit risk or volatility risk.

	Credit risk		Volatility risk	
Horizon h	QIR	cQIR	QIR	cQIR
1	0.14*	0.03	-0.31	-0.16
2	-0.13	0.04	-0.69*	-0.39*
3	-0.12	-0.08	-0.67**	-0.22°
4	-0.35°	-0.17°	-0.65**	-0.25°
5	-0.46**	-0.29°	-0.65°	-0.41*
6	-0.57**	-0.13	-0.72°	-0.36°
7	-0.70**	-0.05	-0.94*	-0.45°
8	-0.74*	-0.24	-1.42*	-0.55°
9	-0.85**	-0.18	-1.78**	-0.06
10	-0.73°	-0.29	-1.52*	-0.22
11	-0.92*	-0.40°	-1.23°	-0.36
12	-0.85°	-0.34	-1.21°	-0.49
13	-1.02*	-0.23	-1.17°	-0.34
14	-1.11**	-0.49°	-0.96°	-0.20
15	-0.96*	-0.55°	-0.90°	0.01
16	-1.17**	-0.51	-0.64	-0.52
17	-1.10*	-0.95*	-0.69	-0.68
18	-0.74°	-0.90°	-0.95	-2.21*
19	-0.68	-1.17**	-0.74	-1.77*
20	-0.78°	-0.58	-1.58°	-1.21°
21	-0.39	-0.87°	-0.50	-0.95
22	-0.35	-0.78°	-0.61	-0.77
23	-0.04	-0.73°	-0.49	-0.36
24	0.27	-0.66°	0.05	-0.55

Table 3: Response of the $\tau = 0.1$ quantile minus the response of the median of Industrial Production (in % pts.) to a shock that increases credit risk or volatility risk by one standard deviation, at horizons $h = 1, \ldots, 24$. QIR: GQLP; cQIR: QLP with the same controls. The marks show that the moving block bootstrap confidence interval (block length of 7, 1,000 bootstrap replications, both quantiles re-estimated in every bootstrap draw) excludes zero at the 68% (°), 90% (*), 95% (**) or 99% (***) confidence level. The intervals are the difference plus or minus z times its bootstrap standard deviation, with $z = 1$, 1.65, 1.96 and 2.58.

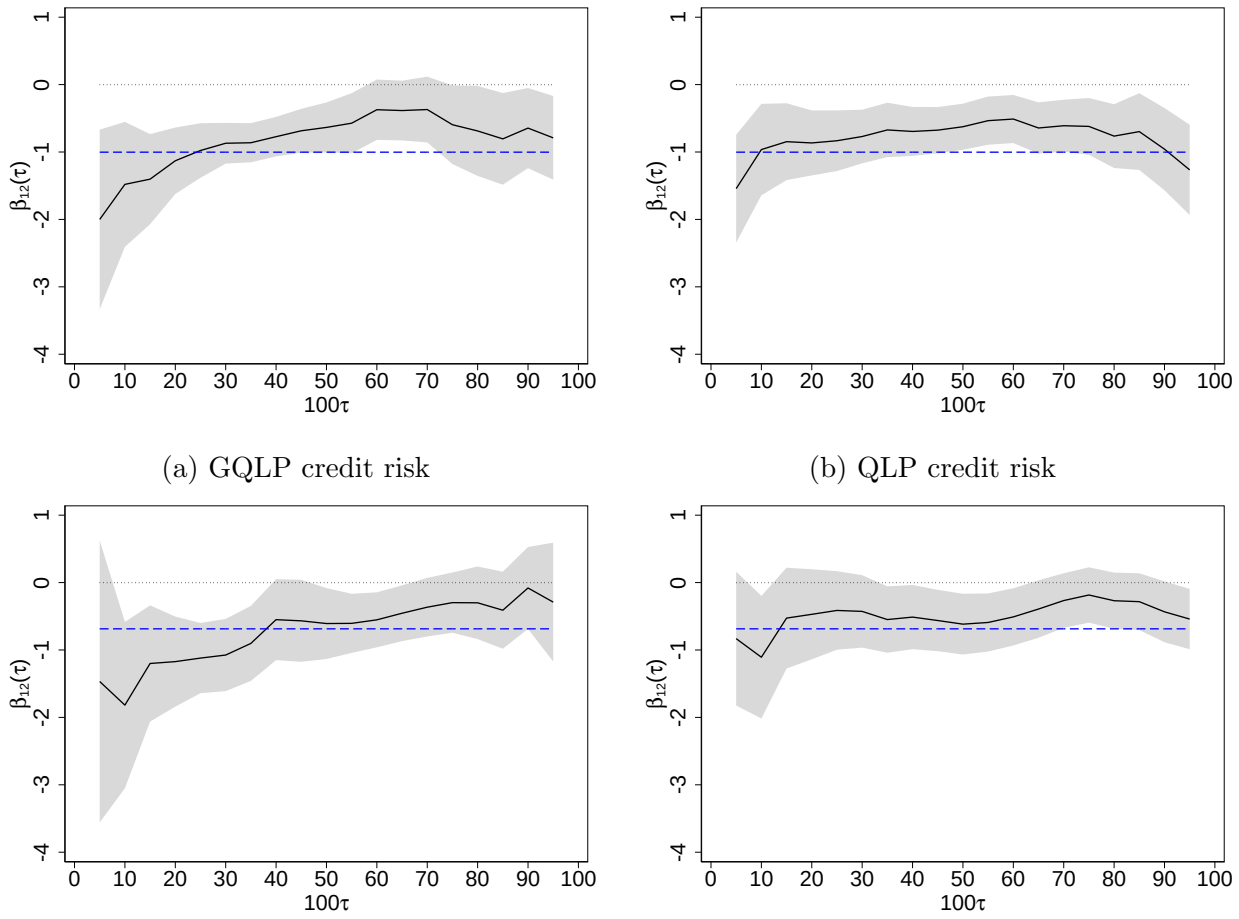
Table 3 tests whether the larger response of the lower tail is statistically significant. It reports the difference between the response of the $\tau = 0.1$ quantile and the response of the median, $\mathsf{QIR}_{0.1}(h) - \mathsf{QIR}_{0.5}(h)$, together with the same difference for QLP, $\mathsf{cQIR}_{0.1}(h) - \mathsf{cQIR}_{0.5}(h)$, at every horizon from one month to two years. For GQLP the difference is negative at 22 of the 24 horizons for credit risk and at 23 for volatility risk, but the intervals are wide. For credit risk I reject the null hypothesis that the difference is zero at the 10% level at 11 horizons between the fifth and the seventeenth month, and at the 5% level at six of them. For volatility risk I reject it at the 10% level at seven horizons between the second and the tenth month, and at the 5% level at three of them. No difference is significant at the 1% level, and beyond 17 months none is significant at the 10% level. The one significant positive difference is in the first month for credit risk, when the median falls by 0.11 percentage points and the lower tail does not move. For QLP the differences are smaller in the first year, and the null hypothesis is rejected at the 10% level at only two horizons for credit risk, both in the second year, and four for volatility risk. In other words, the data support a larger response of the lower tail in the first year after the shock and, for credit risk, until about a year and a half, while the evidence at longer horizons is weak for both estimators.

To highlight the practical implications of using QLP versus GQLP in the context of GaR, I compare the results obtained using both methodologies side by side in Figure 7. Figure 7 plots the response of quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments at the one-year horizon. The estimated shape of the quantile function ($\hat{\beta}_{12}(\tau)$) tells us how much asymmetry there is in the response of different parts of the one-year ahead output growth distribution to a financial shock. A flat line would suggest that all quantiles of the distribution are affected equally meaning that the effect of a financial shock is a location shift of the output growth distribution, implying that quantile analysis is redundant. The fact that the line is upwards sloping means that financial shocks skew the output growth distribution to the left making large negative growth realizations substantially more likely. By comparing the estimates from GQLP versus QLP it is clear that GQLP estimates vary more across quantiles suggesting

larger asymmetry between effects of financial shocks on downside versus median and upside growth scenarios.[19] As the reported results from GQLP and QLP are based on the same timing restrictions and the same data, the difference in the estimates comes mainly from the fact that GQLP captures the effect on unconditional quantiles while QLP captures the effect on conditional on controls quantiles; part of it may also reflect the misspecification of a QLP that is linear in the controls (section 2). This means that volatility and credit risk shocks have a larger negative effect on unconditionally low quantiles of growth than on conditionally low growth quantiles. Therefore, relying on conditional quantile models can understate the importance of these shocks as causes of recessions.

Table 4 tests whether the GQLP and the QLP responses differ. It reports the difference between the two estimates, of the QIR and of the cQIR, at every horizon of the first year and averaged over it, with standard errors from a bootstrap in which both estimators are re-estimated on the same pseudo-samples. In the lower tail the GQLP response is larger in absolute value at 11 of the 12 horizons for credit risk and at all 12 for volatility risk. Averaged over the first year, the difference is -0.29 percentage points for credit risk, with a standard error of 0.19 and a 90% interval from -0.61 to 0.03, and -0.73 for volatility risk, with a standard error of 0.38 and a 90% interval from -1.36 to -0.11. The second is significant at the 10% level ($p = 0.054$) and the first is not ($p = 0.13$). Horizon by horizon, the difference is significant at the 10% level only in the seventh month for credit risk, and at four horizons (3, 4, 9 and 10 months) for volatility risk. At the median the two estimators agree: no difference is significant at the 10% level, and the first-year averages are 0.06 and -0.08. In the upper tail the GQLP response is smaller in absolute value at most horizons, but the first-year averages, 0.11 and 0.21, are not significant at the 10% level. By proposition 1, the two responses coincide in a linear model with independent shocks. The data therefore point to a wedge between the unconditional and the conditional lower-tail responses to volatility risk, which requires a nonlinearity, while for credit risk the first-year difference has the same

[19] Appendix A3.2 reports the QLP responses at every horizon; over the first year they are also less asymmetric than the GQLP responses.



(a) GQLP credit risk

(b) QLP credit risk

(c) GQLP volatility risk

(d) QLP volatility risk

Figure 7: Responses of Industrial Production (in % pts.) to a shock that increases credit risk (top panels) or volatility risk (bottom panels) by one standard deviation, plotted for the one year ahead horizon ($h = 12$). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments using GQLPs (left panels) and QLPs (right panels). Y-axis is the estimated response $\hat{\beta}_{12}(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Blue dashed line reports the response of the mean estimated from conventional local projections. All models include the same variables, assume the same ordering, and have the same lag length specification.

sign but is not statistically significant.

The difference in estimands also changes which periods the estimated lower quantiles single out. Table 5 compares the fitted $\tau = 0.1$ quantile of cumulative growth from GQLP, $\hat{\alpha}_h(0.1) + \hat{\beta}_h(0.1)Y_{j,t}$, with the fitted conditional quantile from QLP, given the treatment and the controls, at $h = 1$ and $h = 6$. By construction, about 10 percent of the months fall below

	Credit risk			Volatility risk		
Horizon h	$\tau = 0.1$	0.5	0.9	$\tau = 0.1$	0.5	0.9
1	0.05	-0.06°	-0.16*	-0.22	-0.06	0.09
	(0.07)	(0.05)	(0.09)	(0.32)	(0.09)	(0.11)
2	-0.14°	0.03	-0.02	-0.30°	-0.00	0.08
	(0.12)	(0.08)	(0.12)	(0.30)	(0.13)	(0.10)
3	-0.06	-0.02	0.03	-0.47*	-0.02	0.27°
	(0.16)	(0.11)	(0.15)	(0.28)	(0.16)	(0.19)
4	-0.19	-0.01	0.20°	-0.53*	-0.12	0.28°
	(0.22)	(0.11)	(0.19)	(0.31)	(0.19)	(0.22)
5	-0.14	0.03	0.12	-0.40	-0.15	0.33*
	(0.22)	(0.15)	(0.22)	(0.40)	(0.23)	(0.19)
6	-0.35°	0.09	0.12	-0.56°	-0.19	0.18
	(0.24)	(0.15)	(0.23)	(0.49)	(0.26)	(0.20)
7	-0.51*	0.14	0.10	-0.62°	-0.13	0.12
	(0.28)	(0.16)	(0.24)	(0.56)	(0.27)	(0.28)
8	-0.34	0.17	0.26	-0.90°	-0.02	0.12
	(0.35)	(0.18)	(0.28)	(0.81)	(0.30)	(0.31)
9	-0.43°	0.24°	0.14	-1.63*	0.10	0.22
	(0.37)	(0.20)	(0.28)	(0.88)	(0.31)	(0.29)
10	-0.37	0.07	-0.01	-1.32*	-0.03	0.20
	(0.38)	(0.20)	(0.31)	(0.73)	(0.32)	(0.30)
11	-0.49°	0.03	0.22	-1.18°	-0.30	0.26
	(0.45)	(0.21)	(0.32)	(0.74)	(0.31)	(0.34)
12	-0.52°	-0.01	0.31	-0.71	0.01	0.35°
	(0.49)	(0.20)	(0.34)	(0.72)	(0.29)	(0.35)
Average, $h = 1, \ldots, 12$	-0.29°	0.06	0.11	-0.73*	-0.08	0.21°
	(0.19)	(0.11)	(0.17)	(0.38)	(0.20)	(0.17)

Table 4: Response of GQLP minus response of QLP with the same controls, $\hat{\beta}_h^{\mathsf{GQLP}}(\tau) - \hat{\beta}_h^{\mathsf{QLP}}(\tau)$, of Industrial Production (in % pts.) to a shock that increases credit risk or volatility risk by one standard deviation, at horizons $h = 1, \ldots, 12$ and averaged over them. Bootstrap standard errors in parentheses. Both estimators are re-estimated, at every quantile and horizon, on the same moving block bootstrap pseudo-samples (block length of 7, 1,000 bootstrap replications). The marks show that the interval, the difference plus or minus z times its bootstrap standard deviation, excludes zero at the 68% (°, $z = 1$), 90% (*, $z = 1.65$), 95% (**, $z = 1.96$) or 99% (***, $z = 2.58$) confidence level.

either. The GQLP quantile is much more closely tied to recessions. In every row of the table, growth falls below it in a larger share of the recession months than below the QLP quantile, by a factor of between 1.7 and 3.5, and a larger share of the months below it are recession months, by a factor of between 1.6 and 3.3. At $h = 1$, growth fell below the GQLP quantile in 52 and 60 percent of the recession months, for credit and volatility risk, against 20 and 17 percent for QLP. At $h = 6$, 87 and 75 percent of the months below the GQLP quantile are

Treatment	h	Recession months	Recession months below the fitted quantile (%)		Months below the quantile in recession (%)	
			GQLP	QLP	GQLP	QLP
Credit risk	1	81	52	20	66	23
	6	116	47	28	87	47
Volatility risk	1	40	60	17	50	15
	6	60	60	35	75	47

Table 5: Recession months and the fitted $\tau = 0.1$ quantile of cumulative growth of Industrial Production, $100\,[\log(IP_{t+h}) - \log(IP_{t-1})]$, from GQLP and from QLP with the same controls, for each treatment and $h \in \{1, 6\}$. Month t is a recession month when the NBER dates place a recession in some month of the window from $t + 1$ to $t + h$, so at $h = 6$ the months just before a recession also count, and there are more recession months than at $h = 1$. Recession months below the fitted quantile: share of the recession months in which the outcome fell below the fitted quantile. Months below the quantile in recession: share of the months below the fitted quantile that are recession months. About 10 percent of all months fall below either fitted quantile.

recession months, against 47 percent for QLP with either treatment. Most months below the conditional quantile are therefore not recessions but months in which growth disappoints given the controls, while low unconditional quantiles map to recessions more closely.

Table 6 compares the estimates obtained from QLP versus GQLP for two horizons $h \in \{6, 12\}$ and three quantiles $\tau \in \{0.1, 0.5, 0.9\}$ using five different lag length specifications: 0, 1, 3, 6 and 12 lags, since applied researchers often include many lags in local projections (Montiel Olea and Plagborg-Møller 2021); with 0 lags only the contemporaneous controls remain. I use this opportunity to point out another potential advantage of GQLP as a tool for quantile impulse response analysis. Since GQLP allows for inclusion of covariates for identification without affecting the interpretation of the coefficient on the treatment variable, it is less sensitive to potentially arbitrary modeling choices such as the choice of how many lags to include. In 9 of the 12 combinations of shock, quantile and horizon in Table 6, including three of the four combinations at $\tau = 0.1$, the estimates vary less across lag lengths with GQLP than with QLP: in these nine combinations the standard deviation of the five QLP estimates is between 1.3 and 9.0 times that of the GQLP estimates. At the 10th quantile, QLP estimates exhibit considerable sensitivity to lag specification, with credit risk effects at the 12-month horizon varying from -1.02 to -1.58 percentage points

and volatility risk effects from -0.06 to -1.20, while the GQLP estimates of the volatility risk effects stay within 0.14 percentage points. At the six-month horizon the QLP estimates range from -0.30 to -0.60 for credit risk and from -0.05 to -0.73 for volatility risk, and the GQLP estimates stay within 0.16 and 0.10 percentage points. The three exceptions are the lower-tail response to credit risk at the 12-month horizon, where the GQLP estimate without lags (-1.07) stands apart from the other four (between -1.36 and -1.73), and the median responses to volatility risk at both horizons. Notably, GQLP tends to identify more negative effects at the lower tail (in 18 of the 20 cells, the exceptions being credit risk at the 12-month horizon with 0 and 12 lags).

Appendix A3.3 reports further checks. First, as in the Monte Carlo designs, the controls can enter GQLP's first stage nonlinearly: with restricted cubic splines of current and lagged industrial production growth and of the lagged shock, or of every control, the point estimates stay inside the baseline 90% bands at every horizon and quantile, and the lower tail still responds much more than the median. Second, the bootstrap bands are not sensitive to the block length: with blocks of 4, 12 or 24 months instead of 7, the half-width of the 90% bands changes on average by at most a fifth. Third, the asymmetry is not a by-product of cumulating the outcome: with three-month growth ending at $t + h$ as the outcome, the lower-tail response averaged over the first year is 3.0 times the median response for credit risk and 5.8 times for volatility risk, against 1.6 for both shocks with QLP (appendix A3.3.5). Fourth, the asymmetry does not come from the pandemic, nor is it specific to industrial production: it remains when the samples end in December 2019, and the $\tau = 0.9$ quantile of the change in the unemployment rate responds more than its median (appendices A3.3.6 and A3.3.7).

To facilitate comparison with previous studies, I focus on the effects at the twelve-month horizon for the median and the 10th quantile. My findings indicate that a one-standard deviation shock to credit or volatility risk lowers the 10th quantile of growth by approximately 1.5 and 1.8 percentage points, respectively, while the losses for median growth are around

		Six-months horizon ($h = 6$)				One-year horizon ($h = 12$)			
		Credit risk		Volatility risk		Credit risk		Volatility risk	
Quantile	Lags	QLP	GQLP	QLP	GQLP	QLP	GQLP	QLP	GQLP
	0	-0.60	-0.76	-0.58	-1.13	-1.33	-1.07	-1.18	-1.68
	1	-0.54	-0.92	-0.73	-1.20	-1.15	-1.73	-0.31	-1.75
	3	-0.32	-0.78	-0.46	-1.10	-1.02	-1.54	-0.86	-1.61
0.1	6	-0.58	-0.78	-0.50	-1.17	-1.07	-1.36	-1.20	-1.75
	12	-0.30	-0.78	-0.05	-1.15	-1.58	-1.54	-0.06	-1.68
	s.d.	0.15	0.06	0.25	0.04	0.23	0.25	0.52	0.06
	ratio	2.3		6.7		0.9		9.0	
	0	-0.12	-0.23	-0.21	-0.39	-0.37	-0.50	-0.39	-0.58
	1	-0.45	-0.37	-0.20	-0.43	-0.59	-0.63	-0.57	-0.67
	3	-0.35	-0.28	-0.37	-0.30	-0.53	-0.58	-0.62	-0.46
0.5	6	-0.39	-0.31	-0.44	-0.38	-0.57	-0.58	-0.52	-0.77
	12	-0.48	-0.31	-0.28	-0.69	-0.79	-0.58	-0.35	-0.82
	s.d.	0.14	0.05	0.10	0.15	0.15	0.05	0.12	0.15
	ratio	2.7		0.7		3.3		0.8	
	0	-0.34	-0.47	-0.30	-0.04	-0.61	-0.69	-0.31	-0.16
	1	-0.38	-0.47	-0.12	0.00	-0.74	-0.61	-0.31	-0.07
	3	-0.42	-0.33	-0.17	0.00	-1.06	-0.53	-0.40	-0.16
0.9	6	-0.62	-0.42	-0.33	-0.01	-1.17	-0.81	-0.28	-0.16
	12	-0.74	-0.40	-0.40	-0.21	-1.21	-1.01	-0.72	-0.20
	s.d.	0.17	0.06	0.12	0.09	0.27	0.19	0.18	0.05
	ratio	2.9		1.3		1.4		3.8	

Table 6: Responses of Industrial Production (in % pts.) to a shock that increases credit risk or volatility risk by one standard deviation, computed for two horizons $h \in \{6, 12\}$ and quantiles $\tau \in \{0.1, 0.5, 0.9\}$. Lags refers to the number of lags included as covariates (0, 1, 3, 6 or 12; the baseline specification has 2 lags; with 0 lags only the contemporaneous controls remain). s.d.: the standard deviation of the five estimates across lag lengths; ratio: the standard deviation of the QLP estimates divided by that of the GQLP estimates. QLP refers to the quantile local projection framework which uses the Koenker and Bassett (1978) estimator. GQLP is my local projections based framework which builds on the generalized quantile regression estimator introduced by Powell (2020). All specifications use the same sample, from the first month with twelve lags of the treatment (February 1974 for credit risk and February 1987 for volatility risk) to August 2026.

0.6 percentage points for both. To the best of my knowledge, this paper is the first to identify the effects of financial shocks on unconditional quantiles of growth, so there are no directly comparable findings. However, a rich literature studies this relationship using conditional quantile models. Adrian et al. (2019), in a quantile regression of one-year-ahead GDP growth on National Financial Conditions Index (NFCI) and current GDP, find losses from a one-standard-deviation increase in NFCI of 1.75 and 0.75 percentage points for the

$\tau = 0.1$ and $\tau = 0.5$ quantiles, respectively. Importantly, the authors do not assign causal interpretation to these estimates as their regression is predictive. Ruzicka (2026) uses QLP (with smoothing) to estimate the effects of NFCI on GDP growth using quarterly data from 1973 to 2015, finding losses of approximately 2.5 and 1.5 percentage points for the $\tau = 0.1$ and $\tau = 0.5$ quantiles, respectively. Loria et al. (2025) study the effects of the EBP on Industrial Production growth using their two-stage methodology. They rescale responses across quantiles such that the median falls by 25 basis points on impact, making direct magnitude comparisons difficult. However, we can compare the ratio of the 10th quantile response to the median response, which Loria et al. (2025) report as 3.7 (averaged over the first year). I find this ratio equals 2.5 for credit risk shocks and 3.7 for volatility risk shocks, computed as the ratio of the average responses over the first twelve months; with QLP the same ratios are 1.4 and 2.1. Loria et al. (2025) examine the effects of various shocks on growth and find similar asymmetries, suggesting a common, non-linear propagation mechanism. This hypothesis is consistent with my findings that volatility risk and credit risk shocks have similar quantile impulse responses in the lower tail despite being only weakly correlated. For the euro area, Chavleishvili and Manganelli (2024) use a quantile vector autoregression to study the effects of shocks to the composite indicator of systemic stress (CISS) on euro area industrial production growth, reporting considerable asymmetry with downside risk losses exceeding median losses by a ratio of approximately 4. Chavleishvili et al. (2021) reach similar conclusions using Bayesian methods.

5 Conclusion

Conventional econometric methods that model the mean impulse responses of growth to financial shocks can underestimate the true importance of financial shocks as causes of recessions. This is widely appreciated by academics and policy-makers alike, which explains why a lot of research effort is devoted to understanding the downside risks to growth.

I offer a new methodology to identify the causal drivers of growth-at-risk. My identification

strategy is based on controls, yet it identifies treatment effects on unconditional quantiles. In my view, the distinction between conditional and unconditional quantiles of growth is important in the context of GaR. Conditionally low growth rates map to periods when the economy under-performs expectations, for example in a favorable macroeconomic climate this would mean high-yet-disappointing growth. On the other hand, unconditionally low growth rates predominantly map to downturns and recessions, and as such are of primary concern for policymakers and academics. My framework allows one to study the latter while using familiar controls-based identification strategies based on timing restrictions.

Understanding the structural drivers of growth vulnerability can help discipline theoretical work and macroprudential policy efforts. My empirical findings show that financial shocks have very large effects on downside risks, larger than their effects on the median and the upper tail of growth. The difference between the lower-tail and the median responses is statistically significant at the 10% level at medium-term horizons (5–17 months for credit risk and 2–10 months for volatility risk), while the evidence at shorter and longer horizons is weaker. Averaged over the first year, the lower-tail response estimated by GQLP exceeds the one estimated by QLP with the same controls by 0.29 percentage points for credit risk and by 0.73 for volatility risk, with bootstrap p-values of 0.13 and 0.054.

Several avenues for future research emerge from this work. First, the methodology could be extended to incorporate instrumental variables identification strategies. Second, future research could extend smoothing techniques to the GQLP framework, similar to how Ruzicka (2026) applies smoothing to QLP, building on the smooth local projection approach of Barnichon and Brownlees (2019). Third, the framework offers promising applications beyond growth-at-risk analysis. One potential application of GQLP could be to study inflation-at-risk, examining how monetary policy shocks differentially affect inflation outcomes in high versus low inflation environments. Such analysis could provide valuable insights into the asymmetric transmission of monetary policy and inform optimal policy design across different inflationary regimes. Another application where GQLP could provide novel insights is studying how the

size of the fiscal multiplier changes depending on whether the government spending shocks occur during expansions versus contractions. Such insights could enhance the effectiveness of fiscal policy in supporting growth by informing the optimal timing of government spending.

References

Adrian, Tobias, Nina Boyarchenko, and Domenico Giannone (2019). "Vulnerable Growth". In: *American Economic Review* 109.4, pp. 1263–89.

Adrian, Tobias, Fernando Duarte, Nellie Liang, and Pawel Zabczyk (2020). "NKV: A New Keynesian Model with Vulnerability". In: *AEA Papers and Proceedings* 110, pp. 470–76.

Adrian, Tobias, Federico Grinberg, Nellie Liang, Sheheryar Malik, and Jie Yu (2022). "The Term Structure of Growth-at-Risk". In: *American Economic Journal: Macroeconomics* 14.3, pp. 283–323.

Angrist, Joshua, Victor Chernozhukov, and Iván Fernández-Val (2006). "Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure". In: *Econometrica* 74.2, pp. 539–563.

Angrist, Joshua D., Òscar Jordà, and Guido M. Kuersteiner (2018). "Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited". In: *Journal of Business & Economic Statistics* 36.3, pp. 371–387.

Angrist, Joshua D. and Guido M. Kuersteiner (2011). "Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score". In: *The Review of Economics and Statistics* 93.3, pp. 725–747. (Visited on 03/08/2023).

Arcones, Miguel A. and Bin Yu (1994). "Central Limit Theorems for Empirical and U-Processes of Stationary Mixing Sequences". In: *Journal of Theoretical Probability* 7.1, pp. 47–71.

Barnichon, Regis and Christian Brownlees (2019). "Impulse response estimation by smooth local projections". In: *Review of Economics and Statistics* 101 (3).

Bloom, Nicholas (2009). "The Impact of Uncertainty Shocks". In: *Econometrica* 77.3, pp. 623–685.

— (2014). "Fluctuations in Uncertainty". In: *Journal of Economic Perspectives* 28.2, 153–76.

Bochmann, Paul, Daniel Dieckelmann, Stephan Fahr, and Josef Ruzicka (2023). "Financial stability considerations in the conduct of monetary policy". In: *ECB Working Paper* 2870.

Brownlees, Christian and André B.M. Souza (2021). "Backtesting global Growth-at-Risk". In: *Journal of Monetary Economics* 118, pp. 312–330.

Brunnermeier, Markus K. and Yuliy Sannikov (2014). "A Macroeconomic Model with a Financial Sector". In: *American Economic Review* 104.2, 379–421.

Campbell, John Y. and Ludger Hentschel (1992). "No News Is Good News: An Asymmetric Model of Changing Volatility in Stock Returns". In: *Journal of Financial Economics* 31.3, pp. 281–318.

Cascaldi-Garcia, Danilo et al. (2023). "What Is Certain about Uncertainty?" In: *Journal of Economic Literature* 61.2, 624–54.

Chavleishvili, Sulkhan, Stephan Fahr, Manfred Kremer, Simone Manganelli, and Bernd Schwaab (2021). "A Risk Management Perspective on Macroprudential Policy". In: *ECB Working Paper Series* 2556.

Chavleishvili, Sulkhan and Simone Manganelli (2024). "Forecasting and stress testing with quantile vector autoregression". In: *Journal of Applied Econometrics* 39.1, pp. 66–85.

Chernozhukov, Victor and Christian Hansen (2005). "An IV Model of Quantile Treatment Effects". In: *Econometrica* 73.1, pp. 245–261. (Visited on 07/29/2025).

— (2008). "Instrumental variable quantile regression: A robust inference approach". In: *Journal of Econometrics* 142, pp. 379–398.

— (2013). "Quantile models with endogeneity". In: *Annual Review of Economics* 5.1, pp. 57–81.

Christensen, B.J. and N.R. Prabhala (1998). "The relation between implied and realized volatility." In: *Journal of Financial Economics* 50.2, pp. 125–150.

Christiano, Lawrence, Martin Eichenbaum, and Charles Evans (1996). "The Effects of Monetary Policy Shocks: Evidence from the Flow of Funds". In: *The Review of Economics and Statistics* 78.1, pp. 16–34.

Chuliá, Helena, Ignacio Garrón, and Jorge M. Uribe (2024). "Daily growth at risk: Financial or real drivers? The answer is not always the same". In: *International Journal of Forecasting* 40.2, pp. 762–776.

Engle, Robert F. and Simone Manganelli (2004). "CAViaR: Conditional autoregressive value at risk by regression quantiles". In: *Journal of Business and Economic Statistics* 22 (4).

Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux (2009). "Unconditional Quantile Regressions". In: *Econometrica* 77.3, pp. 953–973.

French, Kenneth R., G. William Schwert, and Robert F. Stambaugh (1987). "Expected Stock Returns and Volatility". In: *Journal of Financial Economics* 19.1, pp. 3–29.

Galvao, Antonio F., Gabriel Montes-Rojas, and Sung Y. Park (2013). "Quantile Autoregressive Distributed Lag Model with an Application to House Price Returns". In: *Oxford Bulletin of Economics and Statistics* 75 (2), pp. 307–321.

Gertler, Mark, Nobuhiro Kiyotaki, and Andrea Prestipino (2019). "A Macroeconomic Model with Financial Panics". In: *The Review of Economic Studies* 87.1, pp. 240–288.

Gilchrist, Simon and Egon Zakrajšek (2012). "Credit Spreads and Business Cycle Fluctuations". In: *American Economic Review* 102.4, pp. 1692–1720.

Gregory, Karl B., Soumendra N. Lahiri, and Daniel J. Nordman (2018). "A Smooth Block Bootstrap for Quantile Regression with Time Series". In: *The Annals of Statistics* 46.3, pp. 1138–1166.

Hahn, Jinyong and Zhipeng Liao (2021). "Bootstrap Standard Error Estimates and Inference". In: *Econometrica* 89.4, pp. 1963–1977.

Hall, Peter, Joel L. Horowitz, and Bing-Yi Jing (1995). "On Blocking Rules for the Bootstrap with Dependent Data". In: *Biometrika* 82.3, pp. 561–574.

Han, Heejoon, Whayoung Jung, and Ji Hyung Lee (2022). "Estimation and Inference of Quantile Impulse Response Functions by Local Projections: With Applications to VaR Dynamics". In: *Journal of Financial Econometrics* 22.1, pp. 1–29.

He, Zhiguo and Arvind Krishnamurthy (2019). "A Macroeconomic Framework for Quantifying Systemic Risk". In: *American Economic Journal: Macroeconomics* 11.4, pp. 1–37.

Jordá, Ó. (2005). "Estimation and inference of impulse responses by local projections". In: *American Economic Review* 95 (1).

Jorda, Oscar and Alan M. Taylor (2025). "Local Projections". In: *Journal of Economic Literature* 63.1, 59–110.

Jordà, Òscar, Martin Kornejew, Moritz Schularick, and Alan M Taylor (2022). "Zombies at Large? Corporate Debt Overhang and the Macroeconomy". In: *The Review of Financial Studies* 35.10, pp. 4561–4586.

Jung, Whayoung and Ji Hyung Lee (2022). "Quantile Impulse Response Analysis with Applications in Macroeconomics and Finance". In: *Advances in Econometrics*.

Kilian, Lutz and Helmut Lütkepohl (2017). *Structural Vector Autoregressive Analysis*. Themes in Modern Econometrics. Cambridge: Cambridge University Press.

Kiyotaki, Nobuhiro and John Moore (1997). "Credit cycles". In: *Journal of Political Economy* 105 (2).

Koenker, Roger (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.

Koenker, Roger and Gilbert Bassett (1978). "Regression Quantiles". In: *Econometrica* 46 (1).

Kolesár, Michal and Mikkel Plagborg-Møller (2025). "Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly". In: *arXiv preprint*.

Künsch, Hans R. (1989). "The Jackknife and the Bootstrap for General Stationary Observations". In: *The Annals of Statistics* 17.3, pp. 1217–1241.

Lee, Dong Jin, Tae-Hwan Kim, and Paul Mizen (2021). "Impulse response analysis in conditional quantile models with an application to monetary policy". In: *Journal of Economic Dynamics and Control* 127, p. 104102.

Linnemann, Ludger and Roland Winkler (2016). "Estimating nonlinear effects of fiscal policy using quantile regression methods". In: *Oxford Economic Papers* 68 (4).

Loria, Francesca, Christian Matthes, and Donghai Zhang (2025). "Assessing Macroeconomic Tail Risk". In: *The Economic Journal* 135.665, pp. 264–284.

McCracken, Michael W. and Serena Ng (2015). "FRED-MD: A Monthly Database for Macroeconomic Research". In: *Journal of Business & Economic Statistics* 2015-012B. Revision of 2015-012.

Montes-Rojas, Gabriel (2019). "Multivariate Quantile Impulse Response Functions". In: *Journal of Time Series Analysis* 40 (5), pp. 739–752.

Montiel Olea, José Luis and Mikkel Plagborg-Møller (2021). "Local Projection Inference Is Simpler and More Robust Than You Think". In: *Econometrica* 89.4, pp. 1789–1823.

Mumtaz, Haroon and Paolo Surico (2015). "The transmission mechanism in good and bad times". In: *International Economic Review* 56.4, pp. 1237–1260.

Newey, Whitney K. and Daniel McFadden (1994). "Large Sample Estimation and Hypothesis Testing". In: *Handbook of Econometrics*. Ed. by Robert F. Engle and Daniel McFadden. Vol. 4. Elsevier, pp. 2111–2245.

Plagborg-Møller, Mikkel, Lucrezia Reichlin, Giovanni Ricco, and Thomas Hasenzagl (2020). "When is growth at risk?" en. In: *Brookings Pap. Econ. Act.* 2020.1, pp. 167–229.

Plagborg-Møller, Mikkel and Christian K. Wolf (2021). "Local Projections and VARs Estimate the Same Impulse Responses". In: *Econometrica* 89 (2).

Powell, David (2020). "Quantile Treatment Effects in the Presence of Covariates". In: *The Review of Economics and Statistics* 102.5, pp. 994–1005.

Prasad, Ananthakrishnan, Selim Elekdag, Phakawa Jeasakul, Romain Lafarguette, Adrian Alter, Alan Feng, and Changchun Wang (2019). "Growth at Risk: Concept and Application in IMF Country Surveillance". In: *IMF Working Paper* 19.36.

Rambachan, Ashesh and Neil Shephard (2025). "When do common time series estimands have nonparametric causal meaning?" In: *Manuscript, MIT and Harvard University, arXiv:1903.01637v4*.

Ruzicka, Josef (2026). "Quantile local projections: Identification, smooth estimation, and inference". In: *Working paper, Nazarbayev University*.

Sims, Christopher A (1980). "Macroeconomics and Reality". In: *Econometrica* 48.1, pp. 1–48.

van der Vaart, Aad W. and Jon A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. New York: Springer.

Appendix

A1 Econometric framework

A1.1 Proof of lemma 1

Proof. By assumption 5, X_t is a function of $(W_{1:t-1}, W_{-j,t})$, where $W_{-j,t}$ collects the time-t assignments other than $W_{j,t}$, and fixing $Y_{j,t} = y_j$ fixes $W_{j,t} = g_j^{-1}(y_j, X_t)$, a function of the same variables. By assumptions 1 and 4, $Y_{i,t+h}(y_j)$ is therefore a deterministic function of $(W_{1:t-1}, W_{-j,t}, W_{t+1:t+h})$. By assumption 3, $W_{j,t}$ is independent of $(W_{1:t-1}, W_{-j,t}, W_{t+1:t+h})$, hence of the pair $(Y_{i,t+h}(y_j), X_t)$, so $Y_{i,t+h}(y_j) \mid W_{j,t}, X_t \sim Y_{i,t+h}(y_j) \mid X_t$. Finally, by assumption 5, conditioning on $(Y_{j,t}, X_t)$ is the same as conditioning on $(W_{j,t}, X_t)$, because $Y_{j,t} = g_j(W_{j,t}, X_t)$ and $W_{j,t} = g_j^{-1}(Y_{j,t}, X_t)$. □

A1.2 Conditional mean impulse responses

Proposition 2 (Conditional Mean Impulse Response Identification with Unobservable Assignments)**.** *Under Assumptions 1 through 5, if $Y_{i,t+h}(w_j)$ is differentiable in w_j, $g_j^{-1}(y_j, x)$ is differentiable in y_j, and the derivative and the conditional expectation can be exchanged, the causal marginal treatment effect given the controls is identified as:*

$$\frac{\partial}{\partial y_j}\mathbb{E}[Y_{i,t+h} \mid Y_{j,t} = y_j, X_t] = \mathbb{E}\left[\left.\frac{\partial Y_{i,t+h}(w_j)}{\partial w_j}\right|_{w_j = g_j^{-1}(y_j, X_t)} \cdot \frac{\partial g_j^{-1}(y_j, X_t)}{\partial y_j} \mid X_t\right].$$

Proposition 2 is similar to Theorem 10 in Rambachan and Shephard (2025), except that the addition of the partial invertibility assumption 5 makes the interpretation of the identified impulse response more straightforward as it measures the causal effect of an intervention to a single scalar assignment rather than the effect of "simultaneously shifting all assignments from time $t = 1$ to t". The proof is in appendix A1.3. This proposition says that the structural impulse response can be identified without having to observe $W_{j,t}$. Moreover, it takes a

particularly simple form made up of two terms, where the second term $\frac{\partial g_j^{-1}(y_j, X_t)}{\partial y_j}$ does not depend on the horizon of the response nor the dependent variable and is a constant in linear settings. In particular, in a SVAR whose impact matrix has a unit diagonal it is easy to show that $\frac{\partial g_j^{-1}(y_j, X_t)}{\partial y_j} = 1$ and in Local Projections settings it can be normalized to 1 as outlined in Plagborg-Møller and Wolf (2021).

A1.3 Proof of proposition 2

Under assumption 2 the conditional distribution of the assignments given the past admits a strictly positive density, so conditional expectations given $\{Y_{j,t} = y_j, X_t = x\}$ are well defined for every y_j in the support of the treatment; no discreteness of the treatment is required. By assumption 5 the treatment satisfies $Y_{j,t} = g_j(W_{j,t}, X_t)$ with g_j invertible in its first argument, so conditioning on $\{Y_{j,t} = y_j, X_t = x\}$ is the same as conditioning on $\{W_{j,t} = g_j^{-1}(y_j, x), X_t = x\}$. Writing the realized outcome in terms of potential outcomes, $Y_{i,t+h} = Y_{i,t+h}(W_{j,t})$, yields

$$\begin{aligned}\mathbb{E}[Y_{i,t+h} \mid Y_{j,t} = y_j, X_t = x] &= \mathbb{E}\big[Y_{i,t+h}(W_{j,t}) \mid W_{j,t} = g_j^{-1}(y_j, x),\, X_t = x\big] \\ &= \mathbb{E}\big[Y_{i,t+h}(g_j^{-1}(y_j, x)) \mid X_t = x\big],\end{aligned}$$

where the last equality uses the conditional (on X_t) random assignment of lemma 1: once the assignment is fixed at the nonrandom value $w_j = g_j^{-1}(y_j, x)$, the potential outcome $Y_{i,t+h}(w_j)$ is independent of the event $\{W_{j,t} = w_j\}$ conditional on X_t, so the conditioning event carries no further information about it.

Taking partial derivative w.r.t. y_j, exchanging derivative/expectation and applying the

chain rule yields:

$$\frac{\partial}{\partial y_j}\,\mathbb{E}[Y_{i,t+h} \mid Y_{j,t} = y_j, X_t] = \mathbb{E}\Big[\frac{\partial}{\partial y_j} Y_{i,t+h}(g_j^{-1}(y_j, X_t)) \mid X_t\Big]$$
$$= \mathbb{E}\Big[\frac{\partial Y_{i,t+h}(w_j)}{\partial w_j}\Big|_{w_j = g_j^{-1}(y_j, X_t)} \frac{\partial g_j^{-1}(y_j, X_t)}{\partial y_j} \mid X_t\Big].$$

A1.4 Proof of proposition 1

$Y_t \in \mathbb{R}^k$ admits a structural Wold representation

$$Y_t = \sum_{\ell=0}^{\infty} \Psi_\ell W_{t-\ell}, \qquad \sum_{\ell=0}^{\infty} \|\Psi_\ell\| < \infty,$$

where the innovations $\{W_t\}_{t\in\mathbb{Z}}$ are independent over time and have mutually independent components, as i.i.d. $\mathcal{N}(0, I_k)$ innovations do. Fix $i, j \in \{1, \ldots, k\}$ and a horizon $h \geq 0$. Write the i-th component at horizon h as

$$Y_{i,t+h} = [\Psi_h]_{i,j}\, W_{j,t} + R_{i,t+h},$$

where

$$R_{i,t+h} := \sum_{m\neq j} [\Psi_h]_{i,m} W_{m,t} + \sum_{\ell\neq h} \sum_{m=1}^{k} [\Psi_\ell]_{i,m} W_{m,t+h-\ell}.$$

Because $R_{i,t+h}$ is a function of innovations other than $W_{j,t}$, $W_{j,t}$ is independent of $R_{i,t+h}$. Applying the definition of potential outcomes yields:

$$Y_{i,t+h}(w_j) = [\Psi_h]_{i,j}\, w_j + R_{i,t+h}.$$

Mean impulse response (IR):

Taking expectations yields:

$$\mathbb{E}[Y_{i,t+h}(w_j)] = \mathbb{E}[R_{i,t+h} + [\Psi_h]_{i,j} w_j] = \mathbb{E}[R_{i,t+h}] + [\Psi_h]_{i,j}\, w_j.$$

Differentiating w.r.t. w_j yields:

$$\mathsf{IR}(h) = \frac{\partial\, \mathbb{E}[Y_{i,t+h}(w_j)]}{\partial w_j} = [\Psi_h]_{i,j}.$$

Conditional mean impulse response (cIR):

Let X_t be any vector of covariates that is a function of $W_{t-1}, W_{t-2}, \ldots$ and of the components of W_t other than $W_{j,t}$, as assumption 5 requires, so that $W_{j,t}$ is independent of $(R_{i,t+h}, X_t)$. Using the structural Wold representation,

$$\mathbb{E}[Y_{i,t+h}(w_j) \mid X_t = x] = [\Psi_h]_{i,j}\, w_j + \mathbb{E}[R_{i,t+h} \mid X_t = x].$$

Differentiating with respect to w_j gives

$$\mathsf{cIR}(h) = \frac{\partial \mathbb{E}[Y_{i,t+h}(w_j) \mid X_t = x]}{\partial w_j} = [\Psi_h]_{i,j}.$$

Quantile impulse response (QIR):

$$\begin{aligned} q_h(\tau \mid w_j) &\equiv q_{Y_{i,t+h}(w_j)}(\tau \mid W_{j,t} = w_j) \\ &= q_{R_{i,t+h}+[\Psi_h]_{i,j}w_j}(\tau \mid W_{j,t} = w_j) \\ &= q_{R_{i,t+h}}(\tau \mid W_{j,t} = w_j) + [\Psi_h]_{i,j}w_j \\ &= q_{R_{i,t+h}}(\tau) + [\Psi_h]_{i,j}w_j \end{aligned}$$

The first equivalence writes the structural quantile function, the marginal quantile of the potential outcome, as a quantile conditional on $W_{j,t} = w_j$; the two coincide because $Y_{i,t+h}(w_j)$ is independent of $W_{j,t}$. The first equality sign follows from the formula for the potential outcome. The second equality sign follows from the fact that $[\Psi_h]_{i,j}w_j$ is a constant and so it can go outside of the quantile function. The last equality sign follows from the independence

of $W_{j,t}$ w.r.t $R_{i,t+h}$.

Differentiating w.r.t. w_j yields:

$$\mathsf{QIR}_\tau(h) = \frac{\partial\, q_h(\tau \mid W_{j,t} = w_j)}{\partial w_j} = [\Psi_h]_{i,j}.$$

Therefore $\mathsf{QIR}_\tau(h) = [\Psi_h]_{i,j}$ for all $\tau \in (0,1)$.

Conditional quantile impulse response (cQIR):

Let X_t be a function of the other innovations, as above.

$$\begin{aligned} q_h(\tau \mid w_j, x) &\equiv q_{Y_{i,t+h}(w_j)}(\tau \mid W_{j,t} = w_j,\, X_t = x) \\ &= q_{R_{i,t+h}+[\Psi_h]_{i,j}\, w_j}(\tau \mid W_{j,t} = w_j,\, X_t = x) \\ &= q_{R_{i,t+h}}(\tau \mid W_{j,t} = w_j,\, X_t = x) + [\Psi_h]_{i,j}\, w_j \\ &= q_{R_{i,t+h}}(\tau \mid X_t = x) + [\Psi_h]_{i,j}\, w_j \end{aligned}$$

The first equivalence writes the conditional structural quantile function as a quantile conditional on $W_{j,t} = w_j$ and $X_t = x$, which coincide because $Y_{i,t+h}(w_j)$ is independent of $W_{j,t}$ given X_t. The first two equalities follow as in the QIR case. The last equality sign follows from the independence of $W_{j,t}$ w.r.t $R_{i,t+h}$ and X_t, which are functions of the other innovations.

Differentiating w.r.t. w_j yields:

$$\mathsf{cQIR}_\tau(h) = \frac{\partial\, q_h(\tau \mid W_{j,t} = w_j, X_t = x)}{\partial w_j} = [\Psi_h]_{i,j}.$$

Therefore $\mathsf{cQIR}_\tau(h) = [\Psi_h]_{i,j}$ for all $\tau \in (0,1)$.

From the shock to the treatment: The four responses above are derivatives with respect to the shock w_j, whereas equations (1)–(4) are derivatives with respect to the treatment y_j. By the Wold representation, $Y_{j,t} = [\Psi_0]_{j,j} W_{j,t} + r_{j,t}$, where $r_{j,t}$ collects terms in the other innovations and is a function of X_t by assumption 5, so $g_j^{-1}(y_j, X_t) = (y_j - r_{j,t})/[\Psi_0]_{j,j}$, with

$[\Psi_0]_{j,j} \neq 0$. Substituting this shock into the potential outcome gives $Y_{i,t+h}(y_j) = c\, y_j + \tilde{R}_{i,t+h}$, where $c = [\Psi_h]_{i,j}/[\Psi_0]_{j,j}$ and $\tilde{R}_{i,t+h} = R_{i,t+h} - c\, r_{j,t}$ does not depend on y_j. Since $c\, y_j$ is a constant, $\mathbb{E}[Y_{i,t+h}(y_j)] = c\, y_j + \mathbb{E}[\tilde{R}_{i,t+h}]$ and $q_h(\tau \mid y_j) = c\, y_j + q_{\tilde{R}_{i,t+h}}(\tau)$, and the same holds conditional on $X_t = x$. The four responses to the treatment are therefore all equal to $[\Psi_h]_{i,j}/[\Psi_0]_{j,j}$.

This completes the proof that $\mathsf{cIR}(h) = \mathsf{IR}(h) = \mathsf{QIR}_\tau(h) = \mathsf{cQIR}_\tau(h)\ \forall h$ and $\forall \tau$.

A1.5 Proof of theorem 1

I restate theorem 1 of Powell (2020) in my setting. First, I want to show: $\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}, X_t] = \mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t]$. Evaluating the left hand side of the equality yields:

$$\begin{aligned}
\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}, X_t] &= \mathbb{P}[Y_{i,t+h}(Y_{j,t}) \leq q_h(\tau \mid Y_{j,t}) \mid Y_{j,t}, X_t] \\
&= \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid Y_{j,t}, X_t] \\
&= \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid X_t].
\end{aligned}$$

The first equality sign follows from the definition of a potential outcome. The second equality sign comes from the rank similarity assumption 7: given $Y_{j,t}$ and X_t, the conditional probability of the rank event is the same for every treatment value, so it can be evaluated at any fixed y_j in place of the realized $Y_{j,t}$. The third equality sign follows from the conditional (on X_t) independence of $Y_{j,t}$ which follows from assumptions 3, 4 and 5 (lemma 1). Evaluating

the right hand side of the equality yields:

$$\begin{aligned}
\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t}) \mid X_t] &= \mathbb{P}[Y_{i,t+h}(Y_{j,t}) \leq q_h(\tau \mid Y_{j,t}) \mid X_t] \\
&= \int \mathbb{P}[Y_{i,t+h}(Y_{j,t}) \leq q_h(\tau \mid Y_{j,t}) \mid X_t, Y_{j,t}] d\mathbb{P}(Y_{j,t} \mid X_t) \\
&= \int \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid Y_{j,t}, X_t] d\mathbb{P}(Y_{j,t} \mid X_t) \\
&= \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid X_t].
\end{aligned}$$

The first equality follows from the definition of a potential outcome. The third equality follows from the rank similarity assumption 7. The second and fourth equality follow directly from properties of marginal probability functions.

Now I want to show: $\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})] = \tau$.

$$\begin{aligned}
\mathbb{P}[Y_{i,t+h} \leq q_h(\tau \mid Y_{j,t})] &= \int \mathbb{P}[Y_{i,t+h}(Y_{j,t}) \leq q_h(\tau \mid Y_{j,t}) \mid X_t, Y_{j,t}] d\mathbb{P}(X_t, Y_{j,t}) \\
&= \int \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j) \mid X_t, Y_{j,t}] d\mathbb{P}(X_t, Y_{j,t}) \\
&= \mathbb{P}[Y_{i,t+h}(y_j) \leq q_h(\tau \mid y_j)] \\
&= \tau
\end{aligned}$$

The second equality follows from the rank similarity assumption 7. The fourth equality follows from assumption 6.

A1.6 Consistency and asymptotic normality

GQLP is the generalized quantile regression estimator of Powell (2020) with the treatment $Y_{j,t}$ as the instrument and a linear probability model in step 2. Powell shows that the estimator with a probit or logit first stage is consistent and asymptotically normal with independent observations (his theorem 4). Theorem 2 in section 2.4 states the corresponding result for the linear probability model and stationary, weakly dependent data; this section relates its

assumptions to Powell's and proves it.

Fix h and τ. The moments $m_t(\theta, \delta)$ of section 2.4 are Powell's moments (his section III.C), with the normal equations of the linear probability model in place of the score of his probit or logit first stage. The system is exactly identified, so the weighting matrix plays no role, and the GQLP estimates $(\hat{\theta}, \hat{\delta})$ bring the sample mean of m_t as close to zero as its jumps and the step of the grid search allow. The results below are for the exact minimizer of the norm of the sample moment. The profiled search of section 2.4 delivers it: by construction the second block of the sample moment is of order $1/T$ and the third is zero, and the first is a step function of the slope whose jumps are $o_p(T^{-1/2})$ under assumption 8(b). The results apply to a grid estimator whose fine step shrinks faster than $T^{-1/2}$, since theorem 7.2 of Newey and McFadden (1994) requires the estimate to minimize the criterion up to $o_p(T^{-1})$, which in an exactly identified system means that the sample moment is $o_p(T^{-1/2})$ at the estimate. The grid of footnote 9 has a fixed step of 0.005, so the theorem describes an idealized version of the estimator used in this paper; in samples of the size used here the step is small relative to the sampling error.

Assumption 8 of section 2.4 takes the place of Powell's assumption 3. Part (a) replaces independent observations, (b) strengthens his moment condition 3C, (c) gives primitive conditions for his 3D, that the Jacobian $G = \partial\, \mathbb{E}[m_t(\theta, \delta)]/\partial(\theta', \delta')$ at (θ_0, δ_0) exists and is nonsingular (lemma 3), and (d) replaces his $\Sigma = \mathbb{E}[m_t m_t']$ by the long-run variance, because the moments are serially correlated, for example since the h-step outcomes of consecutive periods overlap. His assumption 3B holds trivially for a linear probability model.

Assumption 9 is Powell's assumption $2'$ with the treatment as the instrument and a linear F: (a) and (b) are his $2A'$ and $2B'$, (c) is his $2D'$, (d) his $2E'$ and (e) his $2C'$. Part (c) makes $I_t(\theta)$ continuous in θ with probability one, (d) is the orthogonality condition that allows the linear probability model to be misspecified, and (e) is the identification condition, which requires that the moments of theorem 1 have no other root. It is the global uniqueness condition that GMM estimators of quantile models with endogeneity impose: Powell relates

his 2C′ to assumption R6 of the instrumental variable quantile regression of Chernozhukov and Hansen (2008). Powell's theorems 2 and 3 give sufficient conditions for the moment conditions of theorem 1, which are built with the conditional probability itself, a rank condition when the treatment is discrete and a bounded completeness condition, the analogue of the condition of Chernozhukov and Hansen (2005), when it is continuous. The moments with the linear probability model coincide with those of theorem 1 at every θ when the conditional mean of the treatment is linear in $\tilde{X}_t$, the first case of lemma 2 below. Appendix A2.2.3 checks that the moment conditions have a unique root in the monotone transformation SVAR. The following lemma gives primitive conditions for (d), the condition stated in section 2.

Lemma 2 (First-stage robustness)**.** *Let δ_0 be the coefficient of the linear projection of $I_t(\theta_0)$ on $\tilde{X}_t$. Assumption 9(d) holds if either $\mathbb{E}[Y_{j,t} \mid \tilde{X}_t]$ or $\mathbb{P}(I_t(\theta_0) = 1 \mid \tilde{X}_t)$ is linear in $\tilde{X}_t$.*

Proof. Write $\rho(\tilde{X}_t) = \mathbb{P}(I_t(\theta_0) = 1 \mid \tilde{X}_t)$ and $Y_{j,t} = \tilde{X}_t'\pi + v_t$ for the linear projection of the treatment on $\tilde{X}_t$, so that $\mathbb{E}[\tilde{X}_t v_t] = 0$. Since δ_0 is a linear projection coefficient, $\mathbb{E}[\tilde{X}_t(I_t(\theta_0) - \tilde{X}_t'\delta_0)] = 0$ and, by iterated expectations, $\mathbb{E}[\tilde{X}_t(\rho(\tilde{X}_t) - \tilde{X}_t'\delta_0)] = 0$. Hence $\mathbb{E}\{Y_{j,t}[\rho(\tilde{X}_t) - \tilde{X}_t'\delta_0]\} = \mathbb{E}\{v_t[\rho(\tilde{X}_t) - \tilde{X}_t'\delta_0]\}$. If ρ is linear, the bracket is zero. If $\mathbb{E}[Y_{j,t} \mid \tilde{X}_t]$ is linear, $\mathbb{E}[v_t \mid \tilde{X}_t] = 0$ and the expectation is zero by iterated expectations. □

With a linear probability model the first-stage block of G is $-\mathbb{E}[\tilde{X}_t\tilde{X}_t']$, which does not depend on the parameters, and the nonsingularity of G reduces to that of the 2×2 matrix S of assumption 8(c), with φ_t, π and v_t as in section 2.4.

Lemma 3 (Jacobian)**.** *Suppose that assumption 8(b) holds and that, conditional on $(Y_{j,t}, X_t)$, $Y_{i,t+h}$ has a density that is continuous and bounded by a constant that does not depend on $(Y_{j,t}, X_t)$. Then G exists, and it is nonsingular if and only if $\mathbb{E}[\tilde{X}_t\tilde{X}_t']$ and*

$$S = \begin{pmatrix} \mathbb{E}[\varphi_t v_t] & \mathbb{E}[\varphi_t v_t Y_{j,t}] \\ \mathbb{E}[\varphi_t] & \mathbb{E}[\varphi_t Y_{j,t}] \end{pmatrix}$$

are nonsingular.

Proof. Conditional on $(Y_{j,t}, X_t)$, the expectation of $I_t(\theta)$ is the conditional distribution function of $Y_{i,t+h}$ at $\alpha_h(\tau) + \beta_h(\tau)Y_{j,t}$. Because the conditional density is bounded, the difference quotients of this function in θ are bounded by a constant times $1 + |Y_{j,t}|$, and the weights 1, $Y_{j,t}$ and $\tilde{X}_t$ have finite second moments by assumption 8(b), so by dominated convergence and the continuity of the density the expectations can be differentiated under the integral sign. With the columns ordered as (α, β, δ'),

$$G = \begin{pmatrix} \mathbb{E}[\varphi_t Y_{j,t}] & \mathbb{E}[\varphi_t Y_{j,t}^2] & -\mathbb{E}[Y_{j,t}\tilde{X}_t'] \\ \mathbb{E}[\varphi_t] & \mathbb{E}[\varphi_t Y_{j,t}] & 0 \\ \mathbb{E}[\varphi_t \tilde{X}_t] & \mathbb{E}[\varphi_t Y_{j,t}\tilde{X}_t] & -\mathbb{E}[\tilde{X}_t\tilde{X}_t'] \end{pmatrix}.$$

If $\mathbb{E}[\tilde{X}_t\tilde{X}_t']$ is singular, then $\tilde{X}_t'c = 0$ with probability one for some $c \neq 0$, so $\mathbb{E}[Y_{j,t}\tilde{X}_t']c = 0$ and G maps $(0, 0, c')'$ to zero. If it is nonsingular, subtracting from the first row π' times the last block of rows, which does not change the determinant, replaces $Y_{j,t}$ by v_t in the first two entries of the first row and sets its last block to $-\mathbb{E}[Y_{j,t}\tilde{X}_t'] + \pi'\mathbb{E}[\tilde{X}_t\tilde{X}_t'] = 0$. The first two rows are then $(S, 0)$, and $\det G = \det S \cdot \det\big(-\mathbb{E}[\tilde{X}_t\tilde{X}_t']\big)$. □

With no controls, $\tilde{X}_t$ is a constant, $v_t = Y_{j,t} - \mathbb{E}[Y_{j,t}]$, and S is a row operation on $\mathbb{E}[\varphi_t(1, Y_{j,t})'(1, Y_{j,t})]$, the Jacobian of a linear quantile regression of the outcome on the treatment, so the condition is the usual one for quantile regression. With controls, the treatment is replaced in the first row by the part of it that the controls do not explain, as in section 2.4.

Proof of theorem 2. The proof is that of Powell's theorem 4 (his online appendix A.2), except in the three places where it uses independent observations. First, consistency follows from theorem 2.6 of Newey and McFadden (1994). That theorem is stated for independent observations, but its proof uses independence only in the uniform law of large numbers of their lemma 2.4, which applies the law of large numbers to finitely many functions of the data. For stationary β-mixing data, the ergodic theorem provides it, as Newey and McFadden

(1994) note, and the dominance condition of lemma 2.4, $\mathbb{E}\sup_\Theta \|m_t(\theta,\delta)\| < \infty$, follows from the compactness of Θ and assumption 8(b). Second, the central limit theorem for $T^{-1/2}\sum_t m_t(\theta_0,\delta_0)$ holds for stationary β-mixing data under assumption 8(a) and (b), with the long-run variance Ω in place of Σ: m_t has finite moments of order p by the Cauchy–Schwarz inequality, and the mixing condition $\sum_\ell \ell^{2/(p-2)}\beta_\ell < \infty$, under which the finite-dimensional central limit theorem stated by Arcones and Yu (1994) (their equations 1.4, 1.5 and 1.10) applies, holds because $\nu > p/(p-2)$. Third, Powell obtains stochastic equicontinuity from the Donsker property of the classes of functions in m_t. These classes consist of indicators of half-planes in $(Y_{i,t+h}, Y_{j,t})$ multiplied by 1, $Y_{j,t}$ and the elements of $\tilde{X}_t$, plus terms that are linear in δ, and their envelopes have finite moments of order p by assumption 8(b). For β-mixing data, the uniform central limit theorem of Arcones and Yu (1994) gives the same property for each of these classes, which suffices because the empirical process is linear in the moment function: indicators of half-planes form a Vapnik–Chervonenkis class and the property is preserved by multiplication with a fixed function, the linear terms form a finite-dimensional vector space (van der Vaart and Wellner 1996, lemmas 2.6.15 and 2.6.18), the envelopes are in L^p, and the rate condition of their theorem 2.1, $\ell^{p/(p-2)}(\log \ell)^{2(p-1)/(p-2)}\beta_\ell \to 0$, holds because $\nu > p/(p-2)$. The limit process is continuous with respect to a seminorm that is bounded by a multiple of the L^p norm (their equation 2.9), and $m_t(\theta,\delta)$ is continuous in that norm at (θ_0,δ_0) by assumptions 8(b) and 9(c) and dominated convergence, which gives the stochastic equicontinuity condition of theorem 7.2. Since m_t is a function of the observations (Y_s, X_s) for s between t and $t+h$, it is β-mixing with coefficients $\beta_{\ell-h}$, which satisfy the same conditions. The rest of the proof, which verifies the conditions of theorems 2.6 and 7.2 of Newey and McFadden (1994), is unchanged, with G nonsingular by lemma 3. □

Theorem 2 is for a fixed quantile and horizon. Stacking the moment vectors of several quantiles or horizons gives their joint asymptotic normality under the same assumptions, which is what the tests of differences across quantiles in section 4 require.

A1.7 Bootstrap algorithm

I calculate confidence intervals using a standard moving block bootstrap (Künsch 1989). This procedure preserves the time-dependency by resampling blocks of M consecutive observations instead of resampling individual time points (Kilian and Lütkepohl 2017). After re-estimating the model B times on these pseudo-samples, confidence intervals are constructed as normal-approximation intervals centered at the point estimate, using the bootstrap standard deviation as the standard error. The algorithm proceeds as follows.

1. Bootstrap Setup: Initialize B bootstrap replications and create empty storage for parameter estimates of interest.

2. Block Resampling Loop: For each bootstrap replication $b = 1, \ldots, B$, draw $\lceil T/M \rceil$ starting positions independently and uniformly from $\{1, \ldots, T - M + 1\}$, so that every block lies inside the sample, and construct blocks of M consecutive observations that start at these positions; concatenate the blocks and keep the first T observations to create a pseudo-sample of size T that preserves temporal dependence.

3. Model Estimation: Re-estimate the econometric model on each bootstrap pseudo-sample and extract the parameter estimates of interest. For GQLP, both steps of the estimator are repeated in every pseudo-sample, and the grid search for the slope is centered at the full-sample estimate.

4. Bootstrap Storage: Store the difference between each bootstrap estimate and the original sample estimate: $\hat{\beta}_h(\tau)_b - \hat{\beta}_h(\tau)$ for bootstrap replication b.

5. Confidence Interval Construction: Compute the standard deviation of the B bootstrap estimates and construct normal-based confidence intervals as:

$$\text{CI} = \hat{\beta}_h(\tau) \pm z_{\alpha/2} \times \text{SE}_{\text{bootstrap}}(\hat{\beta}_h(\tau))$$

where

$$\mathrm{SE}_{\mathrm{bootstrap}}(\hat{\beta}_h(\tau)) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\beta}_h(\tau)_b - \bar{\beta}_h^*(\tau)\right)^2}, \qquad \bar{\beta}_h^*(\tau) = \frac{1}{B}\sum_{b=1}^{B}\hat{\beta}_h(\tau)_b,$$

the standard deviation of the bootstrap estimates.

For the difference between the responses of two quantiles (Table 3), and for the difference between the GQLP and QLP responses (Table 4), both are re-estimated on the same pseudo-sample, so the interval accounts for the correlation between the two estimates.

The moving block bootstrap is valid for quantile regression with dependent data (Gregory et al. 2018), and GQLP is an exactly identified GMM estimator with a non-smooth moment, whose consistency and asymptotic normality under stationarity and weak dependence follow from standard arguments (Newey and McFadden 1994), as shown in appendix A1.6. A formal proof that the bootstrap is valid for GQLP is beyond the scope of this paper, and the coverage experiments of section 3.1 provide the evidence. If the moving block bootstrap distribution of $\hat{\beta}_h(\tau)_b - \hat{\beta}_h(\tau)$ is consistent for the asymptotic distribution of theorem 2, then normal intervals built from the bootstrap standard deviation are asymptotically conservative, even when the bootstrap variance does not converge to the asymptotic variance (Hahn and Liao 2021).

In the empirical application I use blocks of $M = 7$ months and $B = 1{,}000$ bootstrap replications. A block length of 7 is of the order of $T^{1/3}$, the rate at which the block length that minimizes the mean squared error of the bootstrap variance grows (Hall et al. 1995): $T^{1/3}$ is about 8.6 in the credit risk sample and 7.9 in the volatility risk sample. Appendix A3.3.4 reports bands with blocks of 4, 12 and 24 months.

A2 Monte Carlo

A2.1 Monte Carlo targets and algorithms

This appendix defines the benchmark of the Monte Carlo study of section 3 and the average QIR, describes the algorithms that compute them by simulation, compares the simulated values with their closed form in the monotone transformation SVAR, and compares the benchmark with the average QIR in the DSGE model.

A2.1.1 The targets

Let $F_h(\cdot \mid y)$ and $f_h(\cdot \mid y)$ be the distribution and density functions of the potential outcome $Y_{i,t+h}(y)$ across periods, so that the SQF is $q_h(\tau \mid y) = F_h^{-1}(\tau \mid y)$, and let f be the density of the treatment $Y_{j,t}$. The QIR of equation (1) is the slope $\partial q_h(\tau \mid y)/\partial y$. When the SQF is not linear, the QIR differs across values of the treatment, so a single response requires a choice of weights. The average QIR

$$\bar{\beta}_h(\tau) = \int \frac{\partial q_h(\tau \mid y)}{\partial y} f(y)\, dy \tag{6}$$

weights the QIR at each value of the treatment by the density f of the treatment, in other words by how often the treatment takes that value. It is the quantile counterpart of an average partial effect and, in this sense, the economic object of interest. However, an estimator that is linear in the treatment cannot recover it when the SQF is curved, even if the treatment is randomly assigned, so it would be too high a bar for linear GQLP. Because the treatment is normal in both models (in the DSGE model up to the truncation of the innovation at ten standard deviations), the average QIR is also the slope of the least squares fit of the SQF with weights $f(y)$: by Stein's lemma, $\int \frac{\partial q_h(\tau|y)}{\partial y} f(y)\, dy = \mathsf{Cov}\big(q_h(\tau \mid Y_{j,t}), Y_{j,t}\big)/\mathsf{Var}(Y_{j,t})$.

The benchmark is what a linear estimator can recover: the population slope $\beta_h^*(\tau)$ of a linear quantile regression of the potential outcome $Y_{i,t+h}(\tilde{y}_j)$ on a treatment value $\tilde{y}_j$

drawn from f independently of the state of the economy and of all future shocks. Formally, $\big(\alpha_h^*(\tau), \beta_h^*(\tau)\big)$ minimizes $\mathbb{E}\big[\rho_\tau\big(Y_{i,t+h}(\tilde{y}_j) - a - b\,\tilde{y}_j\big)\big]$ over (a, b), where $\rho_\tau(u) = u\,(\tau - \mathbf{1}\{u < 0\})$. Its first-order conditions are

$$\int \big[F_h\big(\alpha_h^*(\tau) + \beta_h^*(\tau)\,y \mid y\big) - \tau\big]\,(1, y)'\,f(y)\,dy = 0.$$

Because $F_h(q_h(\tau \mid y) \mid y) = \tau$, the mean value theorem gives $F_h(\alpha_h^*(\tau) + \beta_h^*(\tau)\,y \mid y) - \tau = \bar{f}_h(y)\,\big[\alpha_h^*(\tau) + \beta_h^*(\tau)\,y - q_h(\tau \mid y)\big]$, where

$$\bar{f}_h(y) = \int_0^1 f_h\Big(q_h(\tau \mid y) + u\,\big[\alpha_h^*(\tau) + \beta_h^*(\tau)\,y - q_h(\tau \mid y)\big] \;\Big|\; y\Big)\,du$$

is the density of $Y_{i,t+h}(y)$ averaged between the SQF and the regression line, close to its density at the τ quantile when the line is close to the SQF. The first-order conditions are therefore the normal equations of a least squares fit of the SQF on $(1, y)$ with weights $f(y)\,\bar{f}_h(y)$ (Angrist et al. 2006). As such, the benchmark gives less weight than the average QIR to values of the treatment at which the outcome is dispersed around its τ quantile, and the two coincide when the SQF is linear. When the treatment is randomly assigned, the first-order conditions above are the population moment conditions of a linear quantile regression of the outcome on the treatment and, since the treatment is then independent of the controls, also those of linear GQLP. I therefore compare GQLP and QLP with the benchmark, so that the comparison is mainly about identification rather than functional form (with a curved SQF, linear GQLP applied to data in which the treatment depends on the state need not converge exactly to the benchmark; appendix A2.2.4), and I report the gap between the average QIR and the benchmark as the cost of linearity: the part of the average QIR that a linear specification misses even when identification is perfect.

A2.1.2 The algorithms

The simulation algorithm follows from the definition of potential outcomes and the SQF. Potential outcomes can be generated based on either assignment counterfactuals $Y_{i,t+h}(w_j)$ or counterfactuals relating to a realization of another endogenous variable $Y_{i,t+h}(y_j)$. The first method involves letting the model run for a number of periods before fixing the structural shock in one period to $W_{j,t} = w_j$ for a grid of values, then simulating forward for each w_j grid value and computing the empirical quantiles of $Y_{i,t+h}(w_j)$. This recovers the quantile of the potential outcomes $Y_{i,t+h}(w_j)$ as a function of the fixed assignment w_j, the structural quantile function of the shock. The second method is the same except it requires forcing the endogenous variable to take a value $Y_{j,t} = y_j$, which (to remain consistent with the structural equations) requires choosing the shock value that solves the structural mapping $y_j = g_j(w_j, X_t)$ by setting $w_j = g_j^{-1}(y_j, X_t)$. This second method recovers the SQF $q_h(\tau \mid y_j)$ of equation 1, the object of interest in this paper. Figure 8 illustrates a single repetition of the algorithm.

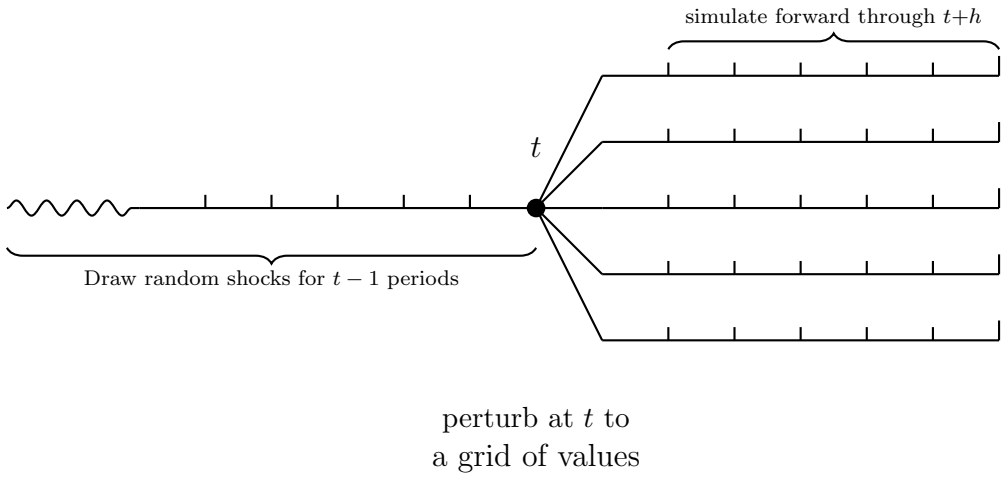


Figure 8: An illustration of the simulation algorithm used to recover the true SQF by simulating a structural model.

The mapping $w_j = g_j^{-1}(y_j, X_t)$ of the second method is particularly simple when g_j is linear, as in both Monte Carlo designs. In particular, if $g_j(W_{j,t}, X_t) = \mu(X_t) + c\,W_{j,t}$ with $c \neq 0$, solving for the shock gives $W_{j,t} = g_j^{-1}(Y_{j,t}, X_t) = c^{-1}[Y_{j,t} - \mu(X_t)]$, so $Y_{j,t} = y_j$ is

equivalent to setting $W_{j,t} = w_j = c^{-1}[y_j - \mu(X_t)]$. Equivalently, compared to an unperturbed draw $\{X_t, W_{j,t}\}$ that produced $Y_{j,t} = \mu(X_t) + c\,W_{j,t}$, the required additive perturbation in the shock is $\Delta = c^{-1}[y_j - Y_{j,t}]$. In both designs, which feature an additive unit-slope ($c = 1$), this further reduces to $\Delta = y_j - Y_{j,t}$, so fixing $Y_{j,t} = y_j$ is exactly the same as perturbing $W_{j,t}$ by adding $\Delta = y_j - Y_{j,t}$. For nonlinear but invertible g_j the mapping need not be an additive shift and must generally be computed pointwise for each X_t. The two methods therefore agree period by period, but not for the distribution of outcomes across periods. Fixing $Y_{j,t} = y_j$ in every period corresponds to a shock $w_j = g_j^{-1}(y_j, X_t)$ that varies with the state, so in a nonlinear model the quantiles of $Y_{i,t+h}(y_j)$ differ from the quantiles of $Y_{i,t+h}(w_j)$ with the shock fixed, and so do their slopes. This is why the benchmarks of section 3 fix the treatment: a quantile regression of the outcome on the structural shock targets the slope of the quantiles of $Y_{i,t+h}(w_j)$ instead.

I compute the targets with two algorithms that follow this logic. Both fix the treatment, by solving for the treatment shock state by state.

Algorithm 1: the SQF and the average QIR.

1. Simulate the model without interventions for many periods. Record the treatment in every period, and the state of the economy in N periods drawn at random.

2. Choose a grid of treatment values $y_1 < \cdots < y_K$ that covers the support of the treatment. For each of the N states and each grid value y_k, find the value of the treatment shock that sets $Y_{j,t} = y_k$ in that state.

3. For each state, draw R paths of future shocks, the same for every grid value, and simulate the model forward h periods from that state and that treatment shock. This gives NR draws of the potential outcome $Y_{i,t+h}(y_k)$ for each grid value.

4. The SQF at y_k is the τ quantile of these draws.

5. The average QIR is $\bar{\beta}_h(\tau) = \sum_{k=1}^{K-1} p_k \,[q_h(\tau \mid y_{k+1}) - q_h(\tau \mid y_k)]/(y_{k+1} - y_k)$, where p_k

is the share of the periods of the simulation in step 1 in which the treatment lies in $[y_k, y_{k+1})$, with the shares beyond the grid added to the first and the last interval.

Algorithm 2: the benchmark.

1. Draw states of the economy from a long simulation without interventions.
2. For each state, draw a treatment value $\tilde{y}_j$ from the distribution of the treatment, independently of the state, and find the value of the treatment shock that sets $Y_{j,t} = \tilde{y}_j$ in that state.
3. Draw future shocks and simulate the model forward h periods. This gives one draw of the potential outcome $Y_{i,t+h}(\tilde{y}_j)$.
4. Run a linear quantile regression of the potential outcomes on the assigned treatment values. Its slope is the benchmark $\beta_h^*(\tau)$.

Up to the discreteness of the grid, Algorithm 2 is a quantile regression through the draws of Algorithm 1 in which each grid value is weighted by its share p_k: in both, the treatment has its own distribution and is independent of the state and of the future shocks.

Implementation. In the SVAR, Algorithm 1 uses $N = 1{,}000{,}000$ states (values of $Z_{i,t}$ in periods drawn from a simulation of 2,000,000 periods), one path of future shocks per state, and a grid from -5.2 to 5.2 in steps of 0.1, about ± 4.6 standard deviations of the treatment. Algorithm 2 uses the periods of a simulation of 500,000 periods as states and their simulated future shocks as paths, and assigns the treatment values by a random permutation of the simulated treatment. Because the latent block is linear and the treatment equation has a unit slope, simulating forward then reduces to $Z_{i,t+s}(\tilde{y}_j) = Z_{i,t+s} + \psi_s(\tilde{y}_j - Y_{j,t})$, where ψ_s is the linear response of the index to the treatment. One such simulation leaves a standard error of 0.015 in the benchmark at $h = 1$ and $\tau = 0.1$, so I average the benchmark over 32 independent simulations; the standard error of the average is at most 0.004 at $\tau = 0.1$ and

0.001 at the median and the upper decile. In the DSGE model, Algorithm 1 uses 4,000 states drawn from the ergodic distribution, excluding the quarters of a run (in which bank net worth is not positive, about 1 percent of quarters), ten paths of future innovations and sunspots per state, common across treatment values, and a grid from -3.5 to 3.5 standard deviations in steps of 0.25 (Figures 3 and 11 show nine of them, from -2 to 2 in steps of 0.5); the shares p_k come from 2,000,000 quarters of capital quality, which is exogenous. Algorithm 2 uses the same states with 25 paths per state, and I average the benchmark over five independent draws of the paths, 500,000 paths in all; the standard error of the average is below 0.006, and the split-half standard error of the average QIR is at most 0.007.

A2.1.3 Closed-form check in the SVAR

In the SVAR the closed form of appendix A2.2.1 gives every target exactly. With $q_h(\tau \mid y) = g\big(\psi_h y + \sqrt{v_h}\,\Phi^{-1}(\tau)\big)$ and $Y_{j,t} \sim N\big(0, \mathsf{Var}(Y_{j,t})\big)$, the average QIR is $\bar{\beta}_h(\tau) = \psi_h\,\mathbb{E}\big[g'\big(\psi_h Y_{j,t} + \sqrt{v_h}\,\Phi^{-1}(\tau)\big)\big]$, and the benchmark solves the first-order conditions of appendix A2.1.1 with $F_h(v \mid y) = \Phi\big((g^{-1}(v) - \psi_h y)/\sqrt{v_h}\big)$. I evaluate both integrals numerically. Table 7 compares these exact values with the simulated ones and with the QIR at $y_j = 0$. The algorithms reproduce the closed form: the simulated average QIR is within 0.002 of the exact one at every horizon and quantile, the simulated SQF is within 0.04 of the closed form, and the simulated benchmark is within 0.008 of the exact one, at most a little more than two of its standard errors. The cost of linearity is small in this model, at most 0.9 percent, because over the range of treatments the model visits the SQF is only mildly concave (section 3.1). For the same reason the QIR at $y_j = 0$ is close to the average QIR.

A2.1.4 The cost of linearity in the DSGE model

Table 8 compares the average QIR with the benchmark in the DSGE model. The lower quantiles of the SQF bend once the adverse move exceeds about one standard deviation, because such moves open a run equilibrium in fragile states (Figure 11). Around the bend

Quantile	Horizon	QIR at $y_j = 0$	Average QIR		Benchmark		Cost of linearity (%)
			Closed form	Algorithm	Closed form	Simulation	
0.1	1	-1.123	-1.110	-1.108	-1.104	-1.105 (0.003)	0.5
	2	-0.852	-0.846	-0.845	-0.845	-0.850 (0.003)	0.2
	4	-0.663	-0.661	-0.660	-0.660	-0.659 (0.004)	0.1
	6	-0.505	-0.504	-0.504	-0.504	-0.502 (0.004)	0.0
	8	-0.384	-0.383	-0.383	-0.383	-0.385 (0.004)	0.0
	10	-0.291	-0.291	-0.291	-0.291	-0.288 (0.003)	0.0
0.5	1	-0.297	-0.304	-0.305	-0.302	-0.303 (0.001)	0.9
	2	-0.223	-0.226	-0.226	-0.225	-0.225 (0.001)	0.5
	4	-0.172	-0.174	-0.174	-0.173	-0.173 (0.001)	0.3
	6	-0.131	-0.131	-0.131	-0.131	-0.131 (0.001)	0.2
	8	-0.099	-0.099	-0.099	-0.099	-0.099 (0.001)	0.1
	10	-0.075	-0.075	-0.075	-0.075	-0.076 (0.001)	0.0
0.9	1	-0.250	-0.250	-0.250	-0.250	-0.250 (0.001)	0.0
	2	-0.188	-0.188	-0.188	-0.188	-0.188 (0.001)	0.0
	4	-0.145	-0.145	-0.145	-0.145	-0.144 (0.001)	0.0
	6	-0.110	-0.110	-0.110	-0.110	-0.109 (0.001)	0.0
	8	-0.083	-0.083	-0.083	-0.083	-0.084 (0.001)	0.0
	10	-0.063	-0.063	-0.063	-0.063	-0.063 (0.001)	0.0

Table 7: Monotone transformation SVAR: the targets of the Monte Carlo study. QIR at $y_j = 0$: the closed-form slope of the SQF at $y_j = 0$, equation (8). Average QIR: equation (6), in closed form and from Algorithm 1. Benchmark: the population slope of a linear quantile regression of the potential outcome on a randomly assigned treatment, in closed form and from Algorithm 2 (average over 32 simulations of 500,000 periods, standard error in parentheses). Cost of linearity: $100\left(\bar{\beta}_h(\tau) - \beta_h^*(\tau)\right)/\bar{\beta}_h(\tau)$, from the closed form.

the potential outcomes are dispersed in the lower tail: on impact, the $\tau = 0.1$ quantile lies 0.74 percent above the $\tau = 0.05$ quantile with no adverse move, and 1.6 to 2.0 percent above it for adverse moves of one to one and a half standard deviations. The benchmark gives these values of the treatment less weight than the average QIR does, so at $\tau = 0.1$ it is 5–11 percent smaller than the average QIR between $h = 0$ and $h = 4$, and at most about 4 percent smaller at longer horizons. At the median and the upper decile, where the SQF is close to linear, the two differ by at most about 2 percent up to $h = 4$, and by at most 4.5 percent at longer horizons, where the responses are small.

Quantile	Horizon	Average QIR	Benchmark	Cost of linearity (%)
0.1	0	-1.352 (0.006)	-1.207 (0.002)	10.7
	1	-1.042 (0.004)	-0.984 (0.001)	5.5
	2	-0.835 (0.001)	-0.795 (0.004)	4.8
	4	-0.558 (0.002)	-0.527 (0.004)	5.5
	8	-0.263 (0.003)	-0.253 (0.001)	3.7
	12	-0.141 (0.003)	-0.139 (0.002)	1.4
0.5	0	-0.968 (0.001)	-0.964 (0.003)	0.5
	1	-0.753 (0.001)	-0.747 (0.003)	0.8
	2	-0.598 (0.000)	-0.592 (0.002)	0.9
	4	-0.397 (0.000)	-0.388 (0.004)	2.2
	8	-0.199 (0.000)	-0.197 (0.003)	1.0
	12	-0.118 (0.002)	-0.112 (0.002)	4.5
0.9	0	-0.887 (0.000)	-0.882 (0.001)	0.6
	1	-0.690 (0.001)	-0.693 (0.001)	-0.4
	2	-0.544 (0.001)	-0.544 (0.005)	-0.0
	4	-0.354 (0.002)	-0.352 (0.001)	0.7
	8	-0.177 (0.001)	-0.177 (0.005)	-0.2
	12	-0.104 (0.000)	-0.108 (0.002)	-3.5

Table 8: GKP model: average QIR and benchmark of the response of log output to a one-standard-deviation adverse move in capital-quality growth, in percent. Average QIR: equation (6), from Algorithm 1 on 29 treatment values between -3.5 and 3.5 standard deviations (split-half standard error in parentheses). Benchmark: from Algorithm 2, the benchmark of Figure 4, averaged over five independent draws of the paths (standard error in parentheses). Cost of linearity: $100\,\big(\bar{\beta}_h(\tau) - \beta^*_h(\tau)\big)/\bar{\beta}_h(\tau)$.

A2.2 Monotone transformation SVAR

A2.2.1 Closed-form responses in the monotone transformation SVAR

Write the first two equations in reduced form as $\eta_t = \Gamma\eta_{t-1} + CW_t$ with $\eta_t = (Z_{i,t}, Y_{j,t})'$, and let $\psi_h \equiv [\Gamma^h]_{12}$ be the linear response of the index to the treatment. Fixing $Y_{j,t} = y_j$ and iterating forward gives the potential outcome

$$Y_{i,t+h}(y_j) = g\big(\psi_h\, y_j + R_{i,t+h}\big), \qquad R_{i,t+h} = [\Gamma^h]_{11} Z_{i,t} + \sum_{s=1}^{h} e_1'\Gamma^{h-s} C\, W_{t+s}, \tag{7}$$

where $e_1 = (1,0)'$. The treatment enters only through the shift $\psi_h y_j$ and the remainder $R_{i,t+h}$ does not involve y_j. By quantile equivariance under the strictly increasing map g, the structural quantile function is $q_h(\tau \mid y_j) = g\big(\psi_h y_j + Q_\tau(R_{i,t+h})\big)$. Because the latent block

Quantile	Horizon	QLP no controls		QLP with controls		GQLP	
		Mean Bias	RMSE	Mean Bias	RMSE	Mean Bias	RMSE
0.1	1	-1.832	1.944	0.613	0.636	0.074	0.502
	2	-1.793	1.913	0.420	0.493	0.118	0.615
	3	-1.629	1.769	0.357	0.486	0.138	0.631
	4	-1.497	1.672	0.267	0.487	0.106	0.663
	5	-1.347	1.540	0.206	0.493	0.111	0.686
	6	-1.185	1.413	0.153	0.515	0.085	0.693
	7	-1.038	1.304	0.139	0.537	0.088	0.705
	8	-0.913	1.202	0.078	0.576	0.022	0.675
	9	-0.803	1.112	0.076	0.578	0.053	0.702
	10	-0.669	1.049	0.079	0.597	0.064	0.718
0.5	1	-0.926	0.969	-0.049	0.105	-0.004	0.102
	2	-0.783	0.830	-0.027	0.116	0.000	0.118
	3	-0.665	0.714	-0.022	0.129	-0.002	0.132
	4	-0.563	0.616	-0.011	0.139	-0.004	0.133
	5	-0.481	0.537	-0.010	0.138	-0.003	0.136
	6	-0.414	0.472	-0.009	0.144	0.001	0.139
	7	-0.353	0.419	-0.005	0.151	-0.003	0.147
	8	-0.299	0.369	-0.002	0.147	-0.002	0.141
	9	-0.253	0.327	0.000	0.147	-0.002	0.145
	10	-0.217	0.297	0.001	0.149	0.002	0.145
0.9	1	-0.649	0.674	-0.035	0.097	-0.009	0.112
	2	-0.554	0.585	-0.005	0.113	-0.008	0.124
	3	-0.480	0.517	-0.001	0.128	0.002	0.136
	4	-0.414	0.457	-0.006	0.138	-0.008	0.142
	5	-0.354	0.404	0.001	0.135	-0.005	0.144
	6	-0.310	0.365	0.003	0.144	-0.007	0.152
	7	-0.258	0.320	0.008	0.147	-0.001	0.154
	8	-0.217	0.289	0.013	0.147	0.007	0.153
	9	-0.188	0.271	0.010	0.155	-0.001	0.162
	10	-0.160	0.247	0.010	0.149	0.000	0.156

Table 9: Monotone transformation SVAR: mean bias and root mean squared error of the three feasible estimators of the QIR (complementing Figure 2 in the main text). The true QIR is the benchmark of Figure 2: the population slope of a linear quantile regression of the potential outcome on a randomly assigned treatment. QLP refers to the quantile local projection estimated with the Koenker and Bassett (1978) estimator; GQLP uses the generalized quantile regression estimator of Powell (2020). Results are from $MC = 1{,}000$ simulation replications, with time-series of length $T = 500$ (after dropping 1,000 burn-in observations).

is linear and Gaussian, $R_{i,t+h} \sim N(0, v_h)$ unconditionally and $R_{i,t+h} \mid X_t \sim N([\Gamma^h]_{11} z_{i,t}, v_h^c)$ conditionally, where $v_h^c = \sum_{s=1}^{h} \|e_1' \Gamma^{h-s} C\|^2$ collects the variance of the shocks that arrive after the treatment and $v_h = [\Gamma^h]_{11}^2 \mathsf{Var}(Z_{i,t}) + v_h^c$ adds to it the variance of the initial condition.

The structural quantile functions and the two impulse responses follow immediately:

$$q_h(\tau \mid y_j) = g\big(\psi_h y_j + \sqrt{v_h}\,\Phi^{-1}(\tau)\big), \qquad \mathsf{QIR}_\tau(h) = \psi_h\, g'\big(\psi_h y_j + \sqrt{v_h}\,\Phi^{-1}(\tau)\big), \tag{8}$$

$$q_h(\tau \mid y_j, x) = g\big(\psi_h y_j + [\Gamma^h]_{11} z_{i,t} + \sqrt{v_h^c}\,\Phi^{-1}(\tau)\big),$$

$$\mathsf{cQIR}_\tau(h) = \psi_h\, g'\big(\psi_h y_j + [\Gamma^h]_{11} z_{i,t} + \sqrt{v_h^c}\,\Phi^{-1}(\tau)\big). \tag{9}$$

As section 3.1 explains, the two responses differ because conditioning on X_t removes the variance contributed by the current state of the economy, so $v_h^c \le v_h$. At the median the two coincide whenever the economy is at its average state, because $\Phi^{-1}(0.5) = 0$. The wedge also closes as h grows, since the initial condition matters less and less for the distribution of the outcome h periods later.

A2.2.2 Rank similarity in the monotone transformation SVAR

I show that the data generating process of section 3.1 satisfies assumption 7 at every horizon $h \ge 1$ and every $\tau \in (0,1)$, with the forwarded outcome $Y_{i,t+h} = g(Z_{i,t+h})$, treatment $Y_{j,t}$ and controls $X_t = \{Y_{i,t}, Y_{j,t-1}, Y_{i,t-1}\}$.

By equation (7), the latent index is additive in the treatment,

$$Z_{i,t+h}(y_j) = \psi_h\, y_j + R_{i,t+h},$$

where $R_{i,t+h}$ does not depend on y_j. Since the amplification map g is strictly increasing ($g'(z) \ge 1 > 0$ by equation (5)), the structural quantile function is that of appendix A2.2.1:

$$q_h(\tau \mid y_j) = g\big(\psi_h\, y_j + Q_\tau(R_{i,t+h})\big),$$

and the rank event does not involve y_j:

$$\begin{aligned}\big\{Y_{i,t+h}(y_j) \le q_h(\tau \mid y_j)\big\} &= \big\{g(\psi_h y_j + R_{i,t+h}) \le g(\psi_h y_j + Q_\tau(R_{i,t+h}))\big\} \\ &= \big\{R_{i,t+h} \le Q_\tau(R_{i,t+h})\big\}.\end{aligned}$$

Consequently, for any conditioning set — and in particular conditional on X_t — the probability of the rank event is the same for all y_j, y_j':

$$\mathbb{P}\big[Y_{i,t+h}(y_j) \le q_h(\tau \mid y_j) \mid X_t\big] = \mathbb{P}\big[R_{i,t+h} \le Q_\tau(R_{i,t+h}) \mid X_t\big],$$

which establishes assumption 7 — indeed, full rank invariance — at every horizon. Note that the conditioning on the observables $X_t = \{Y_{i,t}, Y_{j,t-1}, Y_{i,t-1}\}$ is equivalent to conditioning on $\{Z_{i,t}, Y_{j,t-1}, Z_{i,t-1}\}$ because g is invertible.

The argument uses only two ingredients: additivity of the treatment in the latent linear dynamics and strict monotonicity of the static observation map. Any strictly increasing transformation of a linear Gaussian SVAR variable therefore satisfies rank similarity at all horizons, while featuring τ-dependent QIRs and $\mathsf{QIR}_\tau(h) \neq \mathsf{cQIR}_\tau(h)$ whenever the transformation is nonlinear. Cumulating the transformed outcome across periods would break the single-index structure — a sum of transformed indices is not a monotone function of a single y_j-independent random variable — which is why the outcome in section 3.1 is the forwarded value $Y_{i,t+h}$ rather than a cumulative sum. Appendix A2.4 shows that for the cumulative sum the violation is small and GQLP still recovers the QIR.

A2.2.3 Uniqueness in the monotone transformation SVAR

This appendix checks assumption 9(e) in the monotone transformation SVAR of section 3.1, with the forwarded outcome and a first stage under which the residual of the treatment is the structural shock $W_{j,t}$. With $\lambda > 0$ the SQF is not linear, so assumption 9(b) does not hold, and what is checked is that the linear moment conditions have a unique root, the value

to which linear GQLP converges. A first stage that is linear in $g^{-1}(Y_{i,t})$, $g^{-1}(Y_{i,t-1})$ and $Y_{j,t-1}$ has this property, because the treatment equation is linear in the latent index; it then satisfies assumption 9(d) by lemma 2, and with $\lambda = 0$ it is linear in the controls themselves. By the argument of section 2.4, the moment conditions of theorem 1 then reduce to

$$m_1(a,b) = \mathbb{E}\big[W_{j,t}\,\mathbf{1}\{Y_{i,t+h} \le a + b\,Y_{j,t}\}\big] = 0, \qquad \mathbb{P}(Y_{i,t+h} \le a + b\,Y_{j,t}) = \tau,$$

and given (a,b) the first-stage coefficients δ are unique when the second moment matrix of the first-stage regressors is nonsingular. Uniqueness of the root therefore reduces to uniqueness of (a,b).

By equation (7), $Y_{i,t+h} = g(\psi_h Y_{j,t} + R_{i,t+h})$, and the treatment equation gives $Y_{j,t} = \mu_t + W_{j,t}$, where μ_t depends only on $Z_{i,t}$, $Z_{i,t-1}$ and $Y_{j,t-1}$. Because $Z_{i,t}$ is ordered first, $W_{j,t}$ is independent of $(\mu_t, R_{i,t+h})$. Fix a, b, μ_t and $R_{i,t+h}$, and write the event as $\{\chi(W_{j,t}) \le a\}$, with $\chi(w) = g\big(\psi_h(\mu_t + w) + R_{i,t+h}\big) - b(\mu_t + w)$. Let $\psi_h < 0$, as at every horizon $h = 1, \dots, 10$ of the calibration of Table 1; the case $\psi_h > 0$ is symmetric. With $\lambda > 0$, equation (5) puts the derivative $\chi'(w) = \psi_h g'(\cdot) - b$ strictly between $(1+\lambda)\psi_h - b$ and $\psi_h - b$. If $b \ge \psi_h$, then χ is strictly decreasing, the event is $\{W_{j,t} \ge u\}$ for some $u \in [-\infty, \infty]$ that depends on $(a, b, \mu_t, R_{i,t+h})$, and $\mathbb{E}[W_{j,t}\mathbf{1}\{W_{j,t} \ge u\} \mid \mu_t, R_{i,t+h}] = \phi(u)$, which is positive when u is finite. Since u is finite with positive probability, $m_1(a,b) > 0$ for every a. If $b \le (1+\lambda)\psi_h$, then χ is strictly increasing, the event is $\{W_{j,t} \le u\}$, $\mathbb{E}[W_{j,t}\mathbf{1}\{W_{j,t} \le u\} \mid \mu_t, R_{i,t+h}] = -\phi(u)$, and $m_1(a,b) < 0$ for every a. Every root therefore has $(1+\lambda)\psi_h < b < \psi_h$, an interval that also contains the QIR of equation (8) at every value of the treatment. Given b, the second condition has a unique solution a, because $Y_{i,t+h} - bY_{j,t}$ is continuously distributed with a positive density.

With $\lambda = 0$ the map g is affine and $\chi'(w) = \psi_h - b$, so the same argument gives $m_1(a,b) > 0$ for $b > \psi_h$ and $m_1(a,b) < 0$ for $b < \psi_h$, while at $b = \psi_h$ the event does not involve $W_{j,t}$ and $m_1 = 0$. The root is then unique, and assumption 9(e) holds on any compact Θ that contains it. This is the linear Gaussian SVAR, in which the QIR equals ψ_h at every quantile

(proposition 1).

With $\lambda = 4$, the calibration of section 3.1, the argument confines the roots to an interval but does not rule out several roots inside it. I evaluate $m_1(a(b), b)$, where $a(b)$ solves the second condition, by its sample analogue in one simulation of 10^6 periods, on a grid of 101 values of b that spans the interval, extended by a quarter of $|\psi_h|$ at each end. At every horizon $h = 1, \ldots, 10$ and every $\tau \in \{0.1, 0.5, 0.9\}$ the sample moment changes sign exactly once on the grid, which indicates a single root. At $\tau = 0.9$ the root lies within simulation error of ψ_h, the end of the interval at which the amplification is inactive. At the first horizon the roots are -1.13, -0.30 and -0.25 at $\tau = 0.1$, 0.5 and 0.9, close to the benchmark of Figure 2 (-1.11, -0.30 and -0.25) and to the GQLP estimate in the samples of $T = 10{,}000$ periods of appendix A2.2.4 (-1.13 at $\tau = 0.1$), whose first stage uses splines of the controls instead of the latent index. With $\lambda = 0$ the same computation recovers ψ_h to within 0.004 at $h = 1$, 5 and 10, which measures its simulation error.

A2.2.4 Long samples

In samples of $T = 500$ the lower-tail estimates of GQLP fall short of the benchmark, and so do those of the infeasible quantile regression on the structural shock (Figure 2). Table 10 shows that this is small-sample attenuation. It repeats the experiment of Figure 2 in 500 samples of $T = 10{,}000$ periods, with the same estimators and controls, and reports the mean bias against the benchmark next to the mean bias with $T = 500$. With the long samples the shortfall of GQLP disappears. From the second horizon on, its mean bias is at most 0.01 in absolute value at every quantile, within two Monte Carlo standard errors of zero, against up to 0.14 in the lower tail with $T = 500$. At the first horizon, where the SQF is most curved, its lower-tail estimate settles at -1.13, against a benchmark of -1.11, close to the infeasible regression on the shock (-1.14): with a curved SQF an estimator that is linear in the treatment need not converge exactly to the benchmark. The bias of the two QLPs does not shrink with the sample. QLP with controls still understates the lower-tail response

by 0.61 at the first horizon, 55 percent of the benchmark, and overstates the median and upper-decile responses by 0.05 and 0.04, because it estimates quantiles conditional on the controls. QLP without controls remains far off at every quantile and horizon, because it does not address the endogeneity of the treatment.

		QLP no controls		QLP with controls		GQLP		Shock regression	
Quantile	Horizon	T=500	10,000	500	10,000	500	10,000	500	10,000
0.1	1	-1.832	-1.945	0.613	0.605	0.074	-0.028	0.050	-0.035
	2	-1.793	-1.930	0.420	0.407	0.118	-0.003	0.069	-0.005
	4	-1.497	-1.683	0.267	0.218	0.106	-0.005	0.063	-0.004
	6	-1.185	-1.396	0.153	0.098	0.085	-0.002	0.060	-0.002
	8	-0.913	-1.098	0.078	0.056	0.022	0.003	0.002	0.004
	10	-0.669	-0.853	0.079	0.023	0.064	0.010	0.044	0.009
0.5	1	-0.926	-0.935	-0.049	-0.051	-0.004	0.001	0.006	0.002
	2	-0.783	-0.794	-0.027	-0.035	-0.000	-0.001	0.006	0.000
	4	-0.563	-0.576	-0.011	-0.022	-0.004	-0.001	-0.002	-0.000
	6	-0.414	-0.422	-0.009	-0.012	0.001	-0.001	0.003	-0.001
	8	-0.299	-0.312	-0.002	-0.007	-0.002	0.001	-0.003	0.001
	10	-0.217	-0.233	0.001	-0.003	0.002	0.001	0.001	0.001
0.9	1	-0.649	-0.677	-0.035	-0.037	-0.009	-0.000	0.005	0.002
	2	-0.554	-0.591	-0.005	-0.014	-0.008	0.002	0.003	0.004
	4	-0.414	-0.443	-0.006	-0.001	-0.008	-0.000	-0.003	0.001
	6	-0.310	-0.337	0.003	-0.002	-0.007	-0.003	-0.001	-0.002
	8	-0.217	-0.252	0.013	0.000	0.007	0.000	0.004	0.001
	10	-0.160	-0.194	0.010	-0.003	0.000	-0.002	0.001	-0.002

Table 10: Monotone transformation SVAR: mean bias of the estimators of the QIR against the benchmark, in the 1,000 samples of $T = 500$ periods of Figure 2 and in 500 samples of $T = 10{,}000$ periods. Shock regression: the infeasible quantile regression of the outcome on the structural shock. The Monte Carlo standard errors of the $T = 10{,}000$ biases are at most 0.009 at $\tau = 0.1$ and 0.003 at the median and the upper decile.

A2.3 DSGE model

Quantile	Horizon	QLP no controls		QLP with controls		GQLP	
		Mean Bias	RMSE	Mean Bias	RMSE	Mean Bias	RMSE
Panel A: $T = 320$, GQLP first stage with splines of ξ_{t-1}							
0.1	0	0.816	0.826	0.265	0.265	0.010	0.246
	1	0.660	0.681	0.168	0.219	0.050	0.243
	2	0.518	0.546	0.132	0.212	0.063	0.258
	4	0.361	0.398	0.097	0.215	0.038	0.259
	8	0.174	0.242	0.057	0.212	0.011	0.248
0.5	0	0.581	0.582	0.029	0.031	0.011	0.072
	1	0.465	0.470	0.025	0.074	0.010	0.098
	2	0.383	0.392	0.029	0.094	0.007	0.114
	4	0.263	0.276	0.030	0.105	0.010	0.119
	8	0.135	0.159	0.031	0.119	0.011	0.122
0.9	0	0.520	0.523	-0.050	0.052	-0.016	0.091
	1	0.418	0.431	-0.004	0.088	0.018	0.117
	2	0.334	0.352	0.014	0.107	0.025	0.138
	4	0.230	0.256	0.038	0.131	0.024	0.146
	8	0.123	0.165	0.036	0.141	0.014	0.154
Panel B: $T = 500$, GQLP first stage with splines of all four controls							
0.1	0	0.820	0.825	0.265	0.265	0.011	0.181
	1	0.661	0.674	0.171	0.203	0.064	0.195
	2	0.524	0.541	0.126	0.182	0.083	0.201
	4	0.358	0.381	0.080	0.176	0.071	0.206
	8	0.182	0.219	0.049	0.181	0.053	0.199
0.5	0	0.584	0.585	0.029	0.031	0.005	0.045
	1	0.470	0.474	0.024	0.059	0.016	0.077
	2	0.385	0.391	0.023	0.074	0.023	0.090
	4	0.261	0.268	0.021	0.082	0.029	0.095
	8	0.142	0.156	0.021	0.090	0.034	0.098
0.9	0	0.517	0.519	-0.050	0.051	-0.030	0.071
	1	0.418	0.426	-0.008	0.072	0.015	0.092
	2	0.335	0.346	0.008	0.087	0.030	0.115
	4	0.236	0.252	0.027	0.102	0.041	0.122
	8	0.125	0.155	0.027	0.115	0.028	0.119

Table 11: GKP model: mean bias and root mean squared error of the three feasible estimators of the QIR of log output to a one-standard-deviation adverse move in capital-quality growth (in percent), relative to the benchmark of Figure 4: the population slope of a linear quantile regression of the potential outcome on a randomly assigned treatment. $MC = 500$ replications. Panel A is the design of Figure 4. In panel B the sample is $T = 500$ quarters and GQLP's first stage includes restricted cubic splines of all four controls. QLP with controls always enters the controls $\xi_{t-1}, N_{t-1}, K_{t-1}, \log Y_{i,t-1}$ linearly.

A2.3.1 Bank runs in the model

Figure 9 shows how the economy, affected by a sequence of negative capital-quality shocks, can move from the steady state to a bank-run equilibrium if a sunspot shock occurs. I obtain this figure by re-running the experiment of Figure 2 of Gertler et al. (2019) using the replication package provided by Andrea Prestipino. It uses a typical approach used in theoretical macroeconomic modeling. It is based on a single simulation repetition and reports the impulse responses to a so-called "MIT shock". The "MIT shock" approach, as Kolesár and Plagborg-Møller (2025) describe it, starts the economy at steady state, then hits the economy with a one-off shock (or in this case a sequence of three shocks) while setting all other current and future shocks to zero. Notably, this type of impulse response has no empirical counterpart. They instead define impulse responses as expected counterfactual changes in the outcome, averaging over all other shocks rather than setting them to zero.

A2.3.2 Responses across quantiles

Figure 10 shows the whole cross-quantile profile after one quarter, in the format of the empirical section, with each estimator in its own panel against the benchmark. The benchmark response is -1.16 at $\tau = 0.05$ and rises steadily to -0.66 at $\tau = 0.95$: the adverse move hits the lower tail of output much harder than the upper tail. GQLP traces this profile, staying within 0.05 of the benchmark between $\tau = 0.10$ and $\tau = 0.95$; at $\tau = 0.05$, where only 16 of the 320 observations lie below the quantile, it gives -1.00. QLP with controls is flat at about -0.72 from $\tau = 0.15$ to $\tau = 0.95$, and only its two lowest quantiles bend down. A researcher using it would conclude that the shock shifts most of the distribution of output by the same amount. Between $\tau = 0.05$ and $\tau = 0.95$ its response varies by 0.25 across quantiles, against 0.50 for the benchmark and 0.36 for GQLP.

A2.3.3 Nonlinear specifications of the structural quantile function

Figure 11 shows the structural quantile function after one and four quarters.

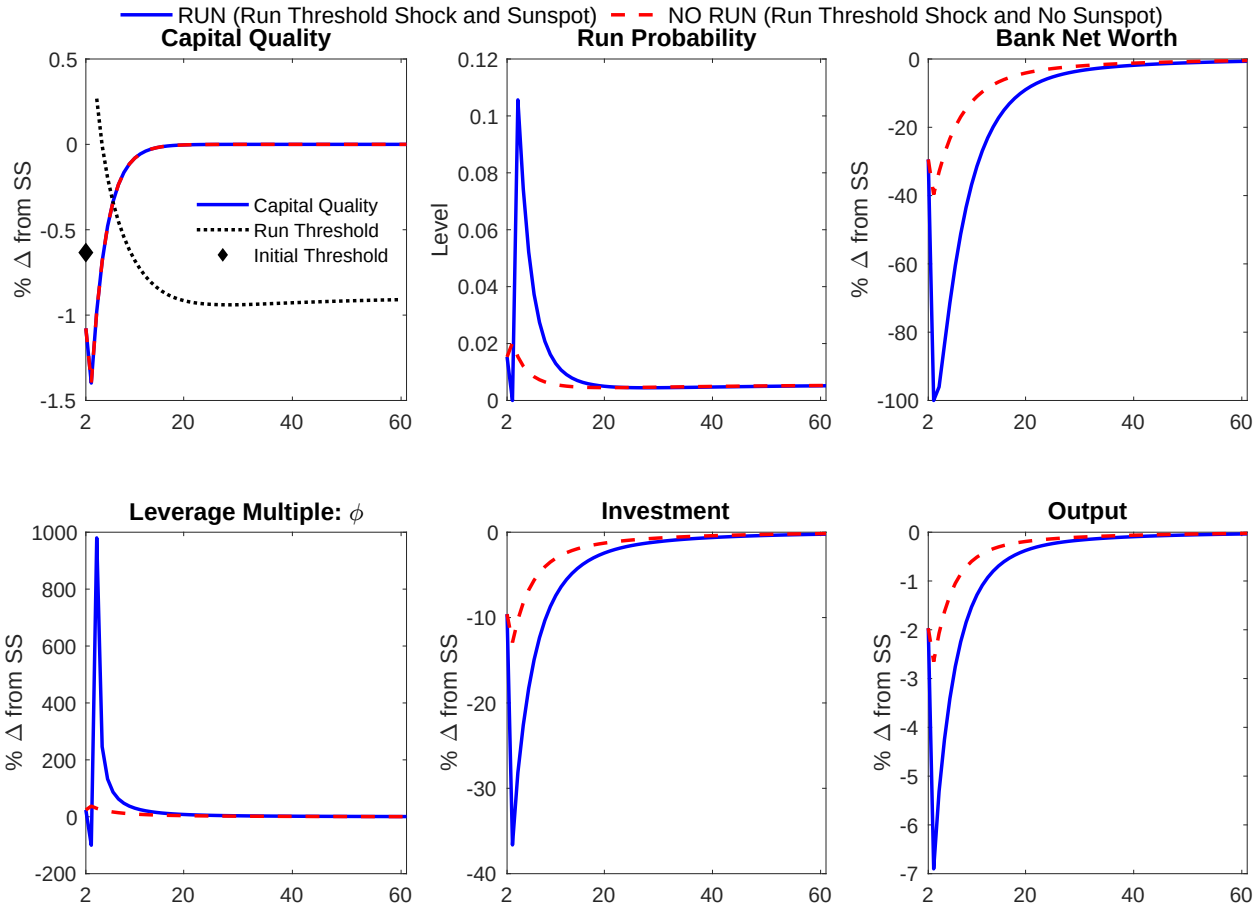
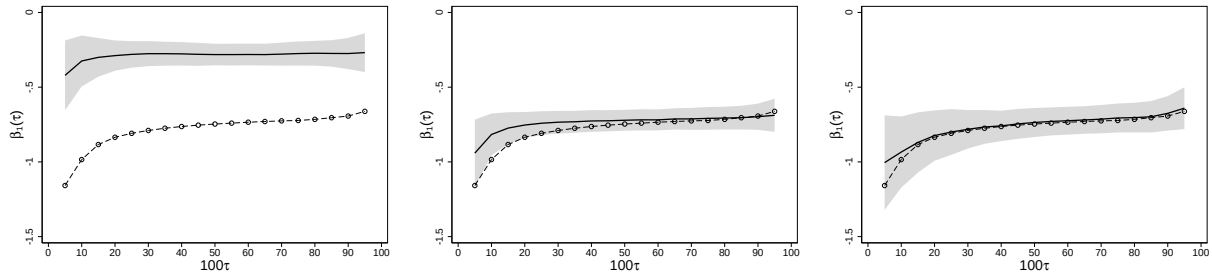


Figure 9: From Gertler et al. (2019), Figure 2, replicated with the authors' code and policy functions. Response of the economy to a sequence of three small negative capital quality shocks combined with a sunspot that triggers a bank run. The plot starts in period 2; the economy is in steady state in period 0, then it experiences the three shocks and no shocks thereafter.

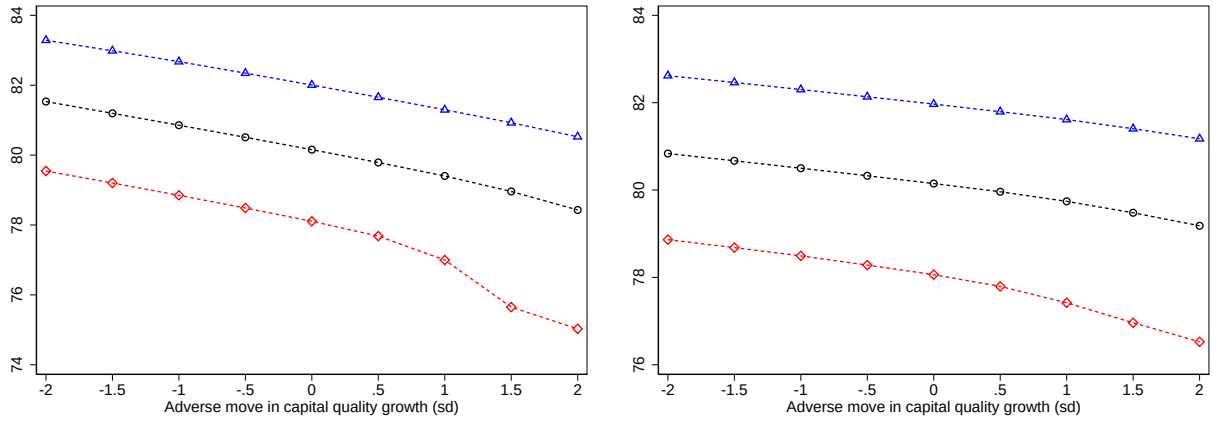
Figure 3 in section 3.2 estimates the structural quantile function after one quarter (left panel of Figure 11) with a linear, a quadratic and a spline specification in the treatment.[20] Section 3.2 discusses the fits. Figure 12 removes the difference in levels by shifting each fitted curve to the structural quantile function at zero; the shifted QLP lower decile still falls by only about 1.85 points, against 2.6–2.7 for GQLP.

[20]The spline is a restricted cubic spline with knots at -1.28, 0 and 1.28 standard deviations of the treatment, the 0.1, 0.5 and 0.9 quantiles of the standard normal distribution; like the quadratic, it has two terms. With two slopes, I minimize the GQLP criterion over a two-dimensional grid: a grid of step 0.1 within ± 2 of the Nelder–Mead estimate, moved if its best point lies on its edge, followed by a grid of step 0.01 around the best point. The QLP fits are evaluated at the sample means of the controls.



QLP without controls, $h = 1$ QLP with controls, $h = 1$ GQLP, $h = 1$

Figure 10: GKP model: responses of log output (in percent) to a one-standard-deviation adverse move in capital-quality growth after one quarter ($h = 1$), for QLP without controls (left), QLP with controls (middle) and GQLP (right). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Solid lines are averages over $MC = 500$ replications of length $T = 320$, and shaded areas are the 68% Monte Carlo bands (the estimate $\pm$ one standard deviation across replications). Dashed lines with circles show the benchmark, the population quantile regression of the potential outcome on a randomly assigned treatment. The three panels share the same vertical axis.

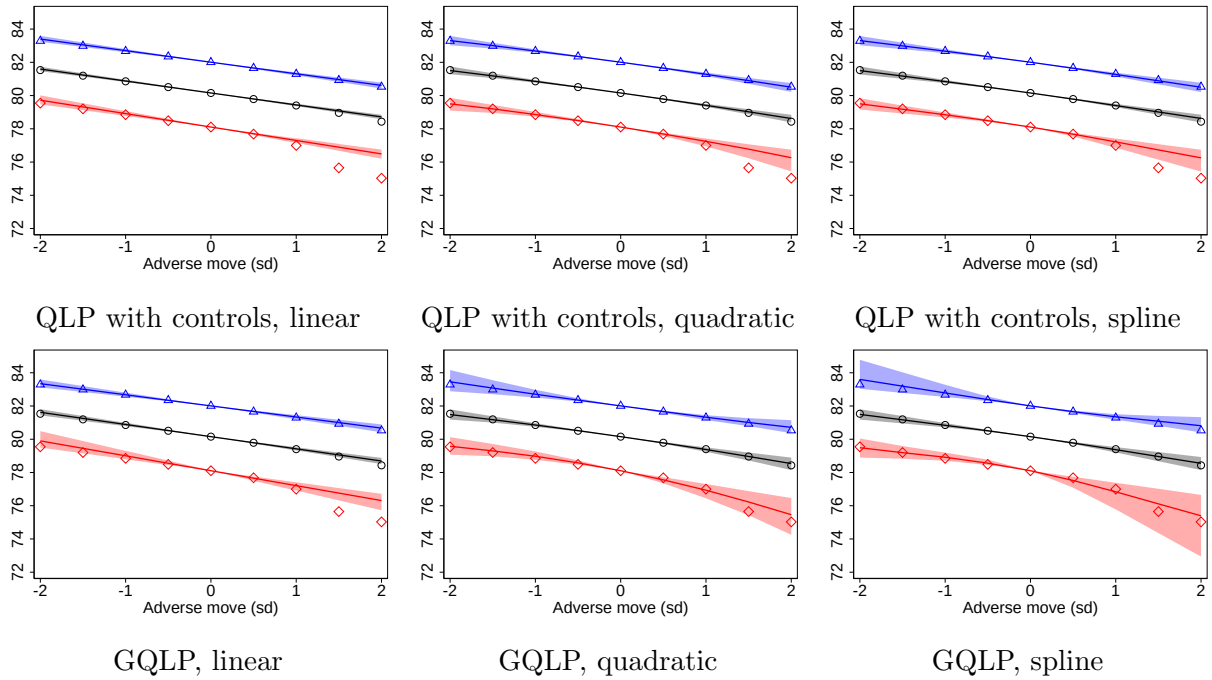


$h = 1$ $h = 4$

Figure 11: GKP model: structural quantile function of $100 \times \log Y_{i,t+h}$ in the treatment, for $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. The horizontal axis is the adverse move in capital-quality growth, in standard deviations. Computed with Algorithm 1 of appendix A2.1.2.

A2.4 Cumulative outcomes

The outcome in the application is cumulative growth, $100\,[\log(IP_{t+h}) - \log(IP_{t-1})]$, and rank similarity is assumed for that outcome. This section checks that GQLP recovers the QIR of a cumulative outcome in the two Monte Carlo designs; appendix A3.3.5 checks that the



QLP with controls, linear | QLP with controls, quadratic | QLP with controls, spline

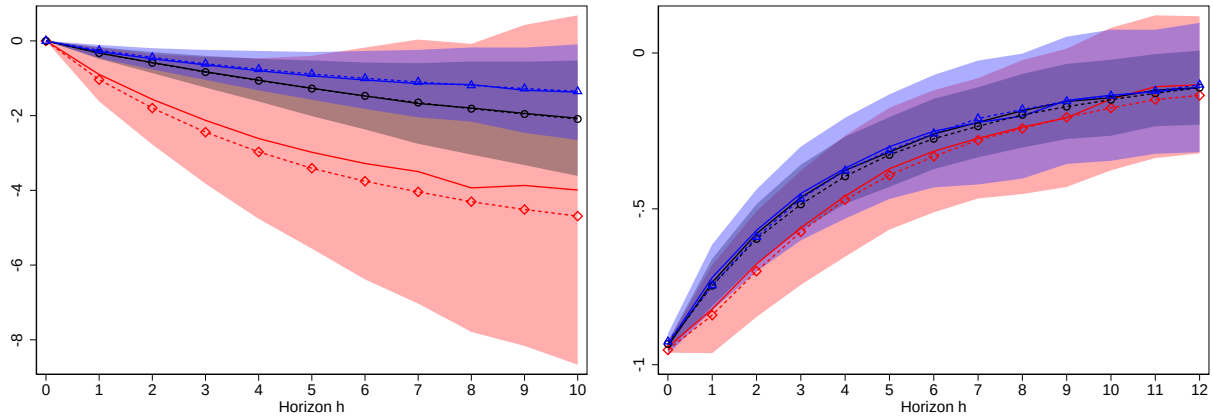
GQLP, linear | GQLP, quadratic | GQLP, spline

Figure 12: GKP model: shape of the estimated structural quantile functions of Figure 3, with linear (left), quadratic (middle) and spline (right) specifications in the treatment, for QLP with controls (top row) and GQLP (bottom row). Each replication's fitted curve is shifted vertically to equal the structural quantile function at zero; solid lines are medians, and shaded areas run from the 0.16 to the 0.84 quantile of the shifted curves across the $MC = 500$ replications. Markers show the structural quantile function, as in Figure 3.

empirical asymmetry does not depend on cumulating.

In the monotone transformation SVAR, rank similarity holds exactly for the forwarded outcome but not for the cumulative sum $\sum_{s=0}^{h} Y_{i,t+s}$ (appendix A2.2.2). The violation is small: at every horizon the Spearman rank correlation between the potential outcomes at treatment values two standard deviations below and above the mean is at least 0.998, and the probability of the rank event conditional on the state varies by at most 1.5 percentage points across treatment values (0.7 points at $\tau = 0.1$), both at the average state and at a low state. Panel (a) of Figure 13 repeats the experiment of Figure 2 with the cumulative outcome, against the benchmark computed in the same way. GQLP stays within 15 percent of the benchmark at $\tau = 0.1$ and within 2 percent at the median at every horizon (13 percent at $\tau = 0.9$ at the first horizon), while QLP with controls understates the lower-tail

response by 30–53 percent, as it does with the forwarded outcome. Panel (b) repeats the DSGE experiment of Figure 4 with the outcome $\log Y_{i,t+h} - \log Y_{i,t-1}$. With this outcome the benchmark response is nearly symmetric on impact (-0.95 at $\tau = 0.1$ against -0.93 at the median) and asymmetric from the first quarter on (-0.84 against -0.75); GQLP is within 0.04 of the benchmark at every horizon and quantile.

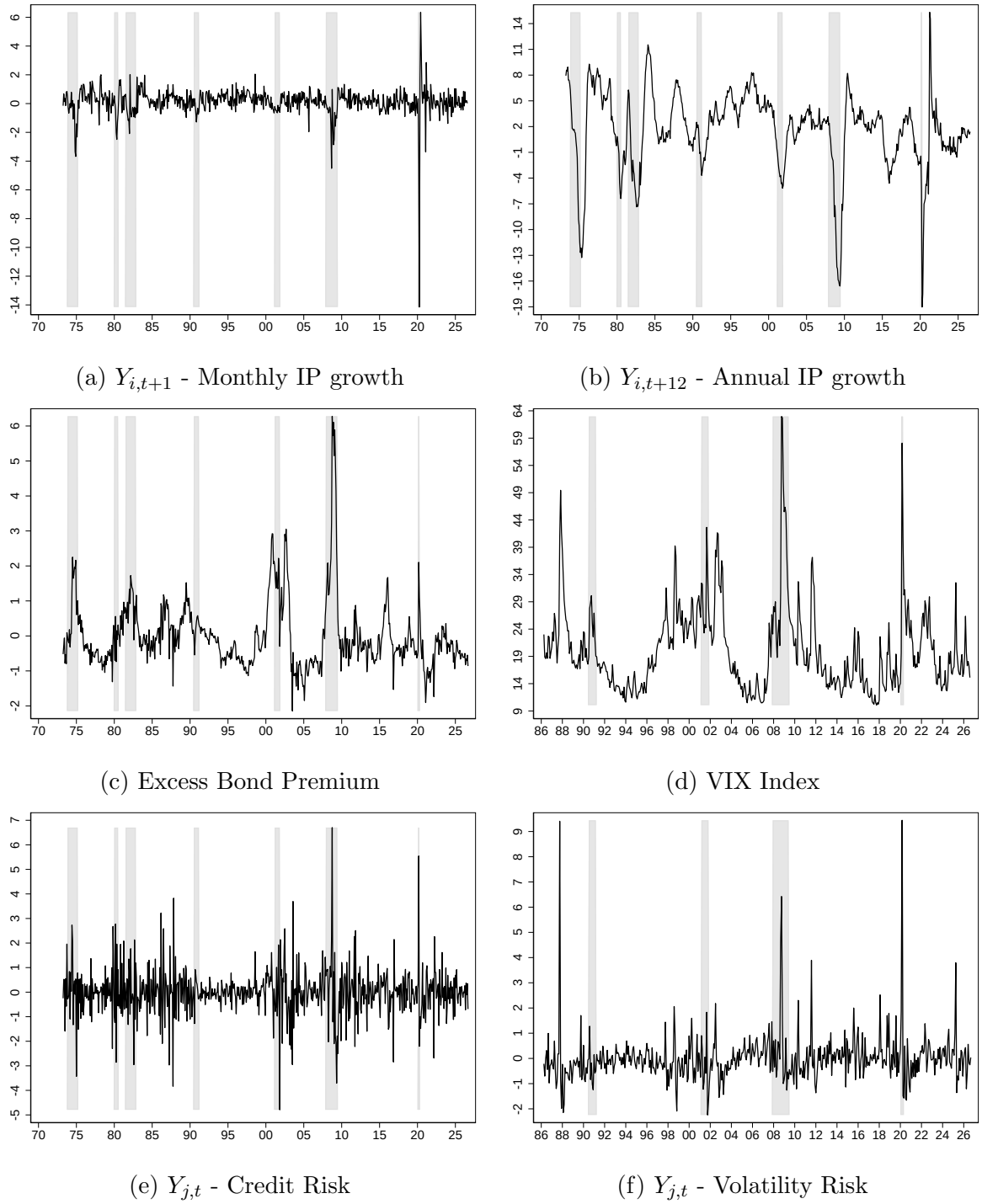


(a) Monotone transformation SVAR, GQLP

(b) GKP model, GQLP

Figure 13: Cumulative outcomes. (a) Monotone transformation SVAR with the cumulative outcome $\sum_{s=0}^{h} Y_{i,t+s}$: GQLP (lines, with 68% Monte Carlo bands) against the benchmark (markers), as in Figure 2. (b) GKP model with the outcome $\log Y_{i,t+h} - \log Y_{i,t-1}$, in percent, as in Figure 4. Quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$.

A3 Empirical results



(a) $Y_{i,t+1}$ - Monthly IP growth

(b) $Y_{i,t+12}$ - Annual IP growth

(c) Excess Bond Premium

(d) VIX Index

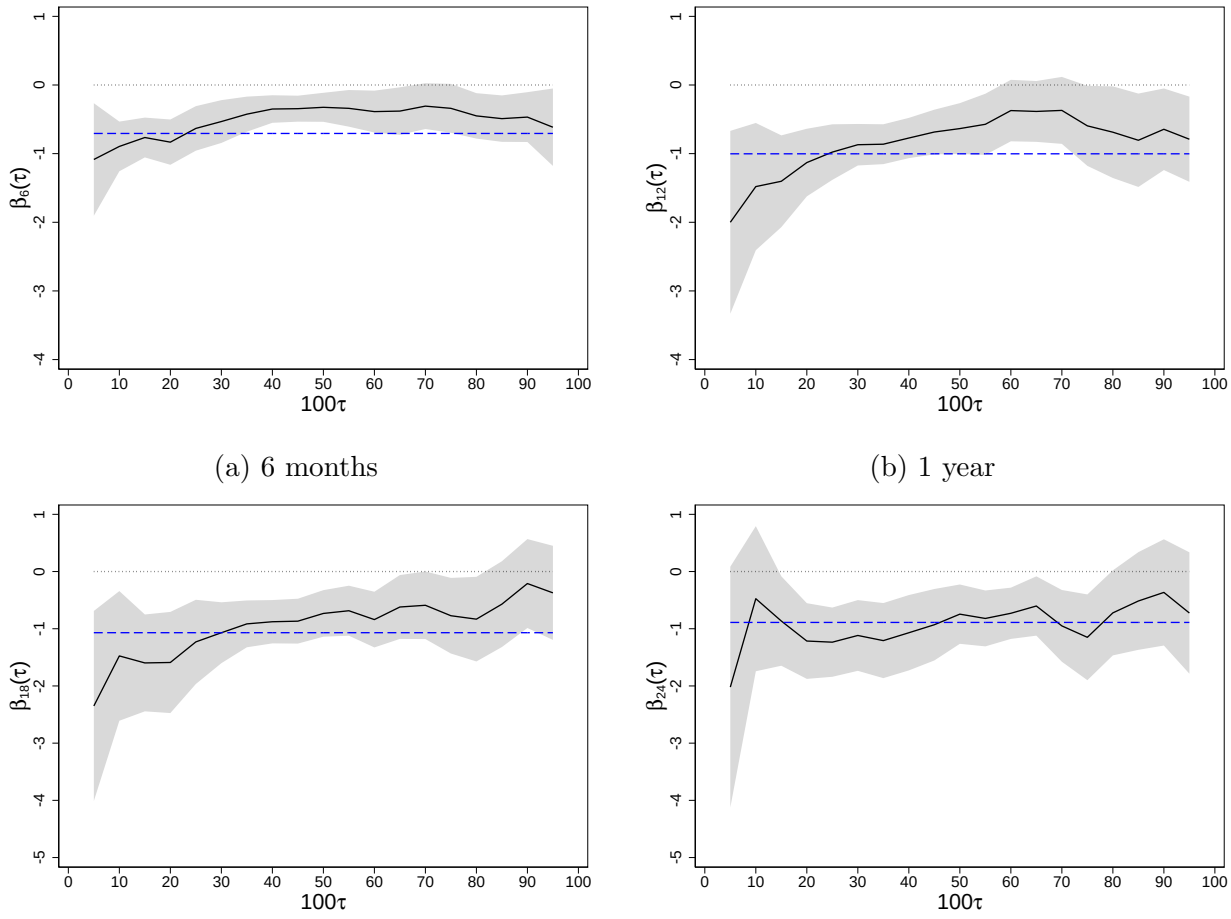
(e) $Y_{j,t}$ - Credit Risk

(f) $Y_{j,t}$ - Volatility Risk

Figure 14: Monthly time series from April 1973 (April 1986 for the VIX and volatility risk) to August 2026 (September 2026 for the excess bond premium and volatility risk). Gray bands indicate NBER recession dates. The series in the bottom panels have been normalized.

A3.1 Responses across quantiles

Figures 15 and 16 report the GQLP responses of industrial production for quantiles from $\tau = 0.05$ to $\tau = 0.95$ at four horizons, discussed in section 4.



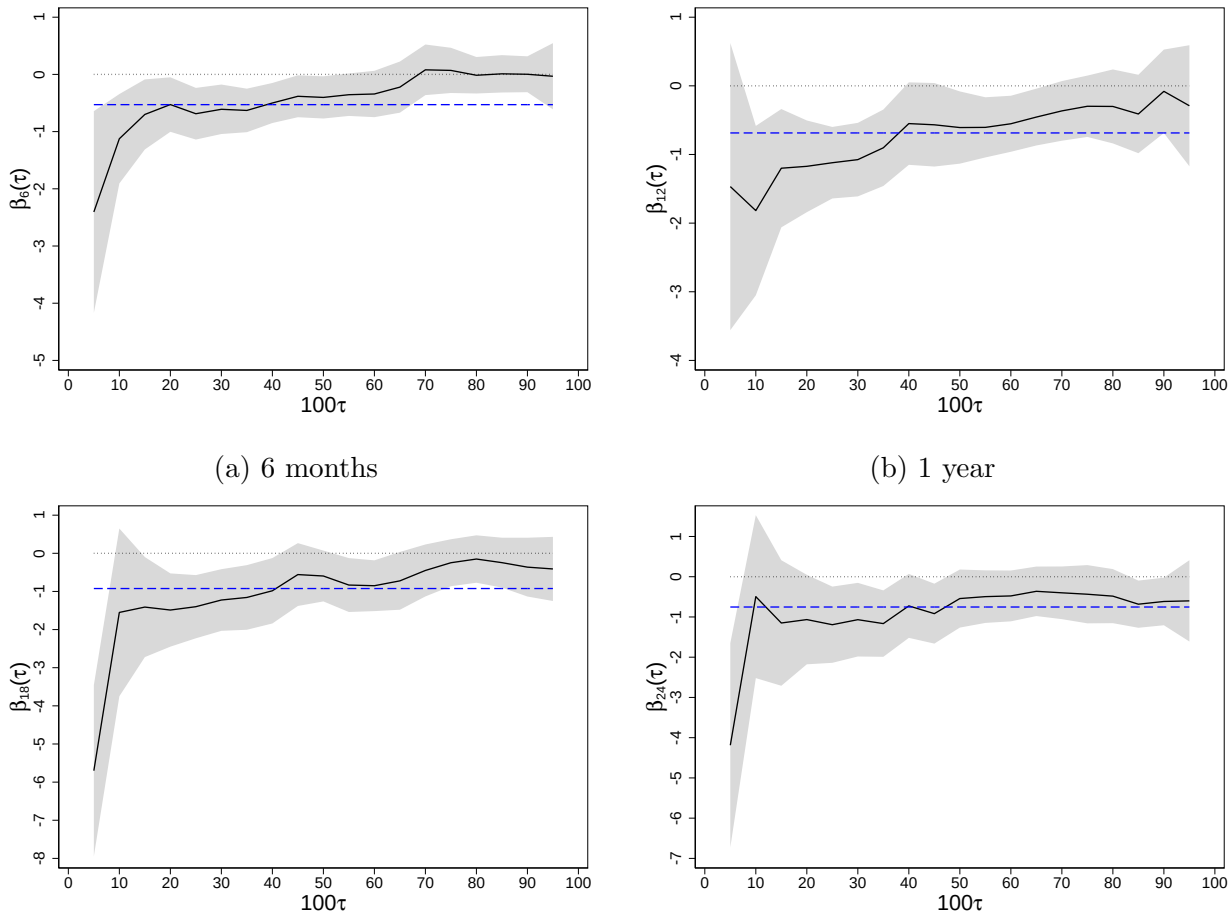
(a) 6 months

(b) 1 year

(c) 18 months

(d) 2 years

Figure 15: Responses of Industrial Production (in % pts.) to a shock that increases credit risk by one standard deviation, plotted for four horizons $h \in \{6, 12, 18, 24\}$ (panels from top left to bottom right). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Blue dashed line reports the response of the mean estimated from conventional local projections.



(a) 6 months

(b) 1 year

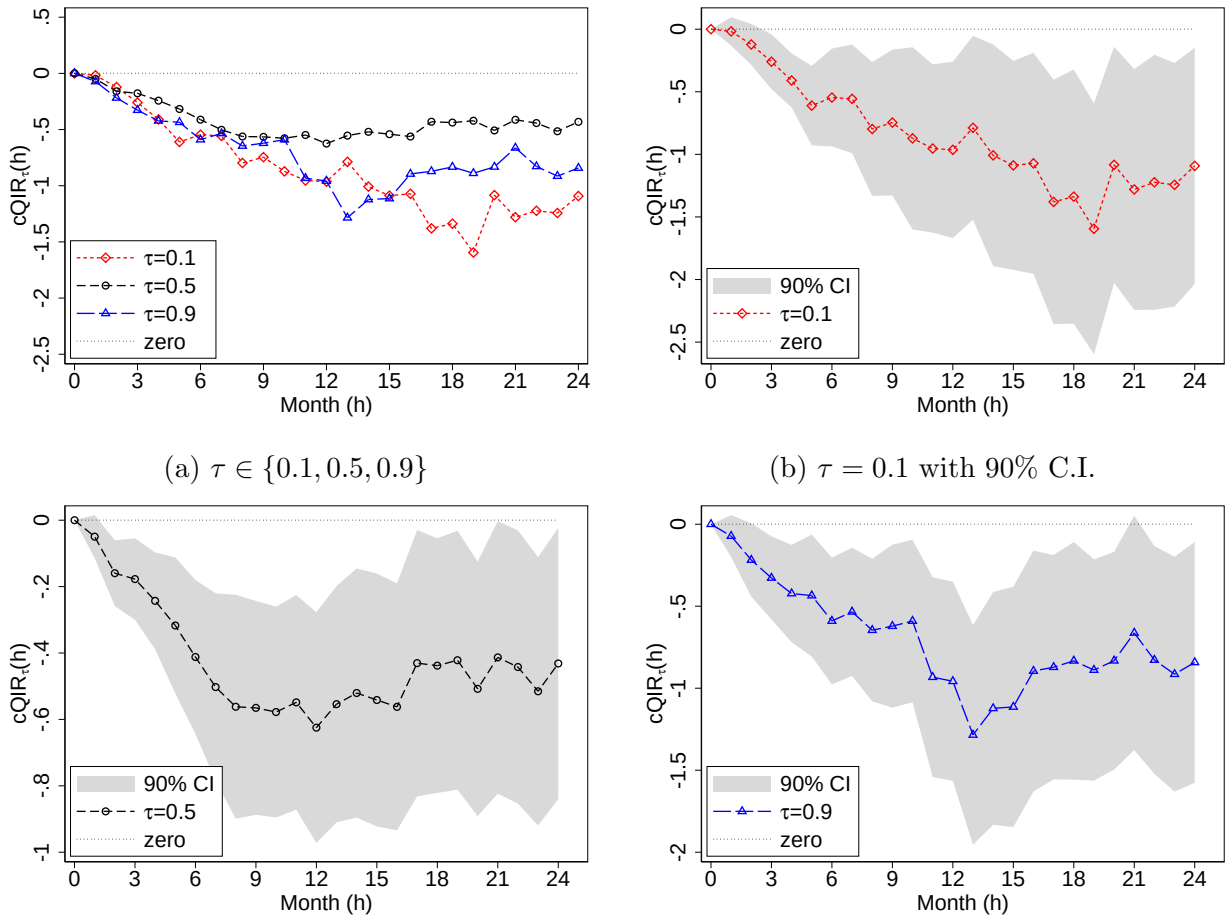
(c) 18 months

(d) 2 years

Figure 16: Responses of Industrial Production (in % pts.) to a shock that increases volatility risk by one standard deviation, plotted for four horizons $h \in \{6, 12, 18, 24\}$ (panels from top left to bottom right). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Blue dashed line reports the response of the mean estimated from conventional local projections.

A3.2 Conditional quantile impulse responses

Figures 17 and 18 report the conditional quantile impulse responses, estimated by QLP with the baseline controls, timing restrictions and samples. They are the counterparts of Figures 5 and 6. The QLP responses are less asymmetric. Over the first year the response of the $\tau = 0.1$ quantile averages -0.57 percentage points for credit risk and -0.62 for volatility risk, 1.4 and 2.1 times the median response, against 2.5 and 3.7 times with GQLP. For credit risk



(a) $\tau \in \{0.1, 0.5, 0.9\}$

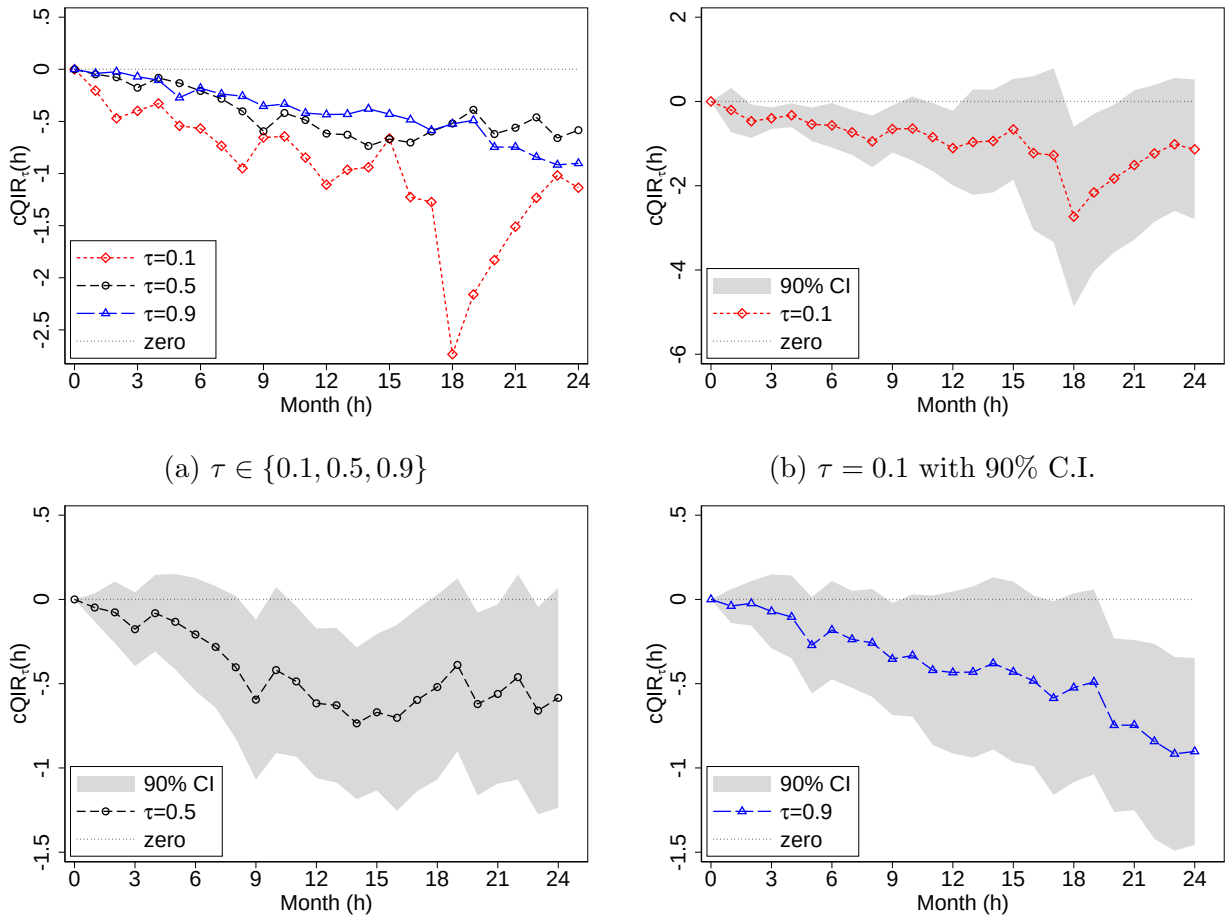
(b) $\tau = 0.1$ with 90% C.I.

(c) $\tau = 0.5$ with 90% C.I.

(d) $\tau = 0.9$ with 90% C.I.

Figure 17: Cumulative response of Industrial Production (in % pts.) from a shock that increases credit risk by one standard deviation, estimated by QLP with the baseline controls (the cQIR), plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the conditional quantile impulse response. Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Note that the impact response (horizon $h = 0$) is by assumption zero, given my timing restrictions.

the $\tau = 0.9$ quantile responds almost as much as the $\tau = 0.1$ quantile (its first-year average is -0.53), so the shock mostly shifts the conditional distribution down. The lower-tail responses reach their troughs later than with GQLP, at -1.59 percentage points after 19 months for credit risk and -2.73 after 18 months for volatility risk. Their 90% bands exclude zero from the third month onwards for credit risk, and for volatility risk at most horizons of the first year and from the 18th to the 20th month.



(a) $\tau \in \{0.1, 0.5, 0.9\}$

(b) $\tau = 0.1$ with 90% C.I.

(c) $\tau = 0.5$ with 90% C.I.

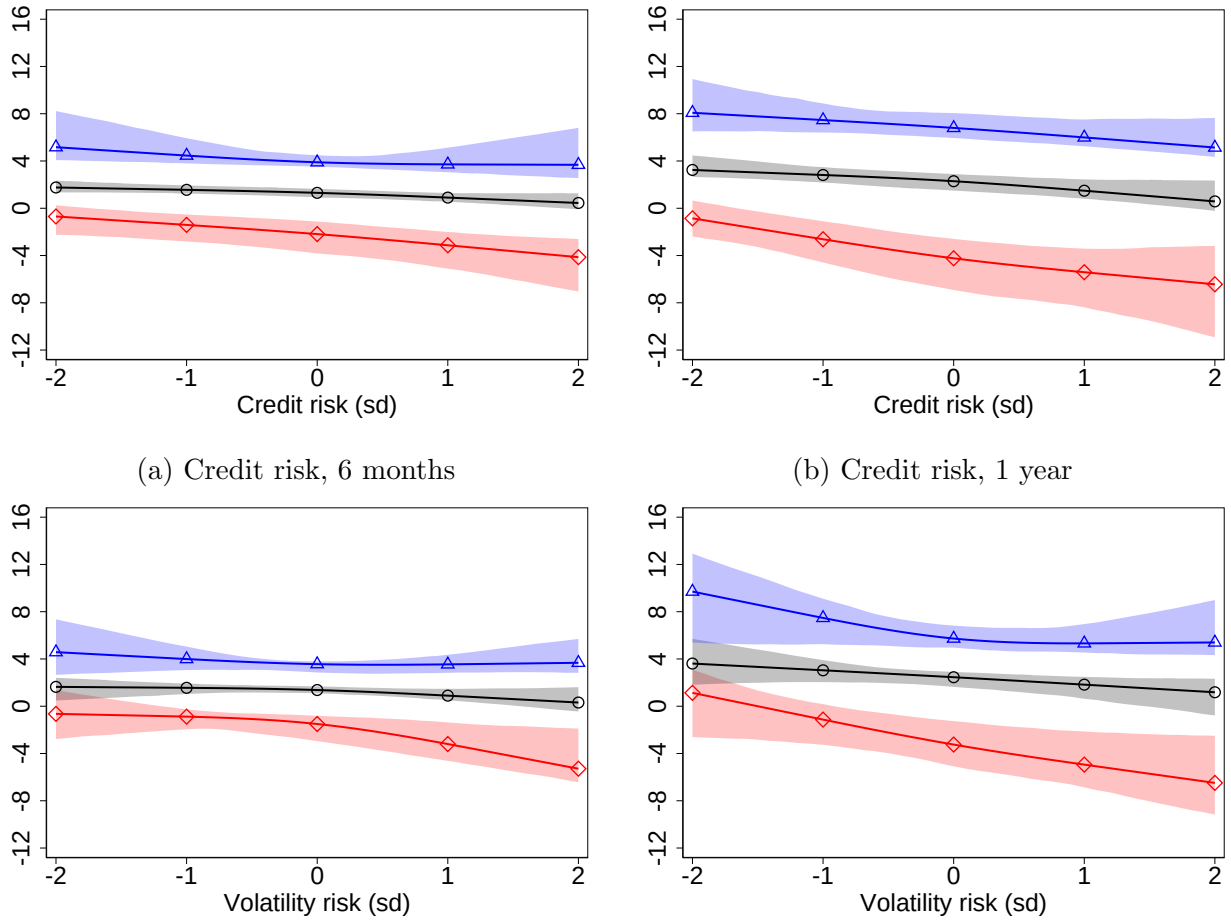
(d) $\tau = 0.9$ with 90% C.I.

Figure 18: Cumulative response of Industrial Production (in % pts.) from a shock that increases volatility risk by one standard deviation, estimated by QLP with the baseline controls (the cQIR), plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the conditional quantile impulse response. Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Note that the impact response (horizon $h = 0$) is by assumption zero, given my timing restrictions.

A3.3 Robustness

A3.3.1 Non-linear functional forms of the SQF

All the empirical results reported in the main text are from a specification that is linear in the treatment. As in the DSGE exercise of Figure 3, GQLP can also estimate the structural quantile function $q_h(\tau \mid y_j)$ itself with a nonlinear specification. Figure 19 reports GQLP



(a) Credit risk, 6 months

(b) Credit risk, 1 year

(c) Volatility risk, 6 months

(d) Volatility risk, 1 year

Figure 19: GQLP estimates of the structural quantile function $q_h(\tau \mid y_j)$ of the cumulative growth of Industrial Production (in % pts.) with a spline specification in credit risk (top row) and volatility risk (bottom row), at the 6-month (left) and 1-year (right) horizons, for quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. The horizontal axis is the treatment, in standard deviations. Solid lines are the fitted structural quantile functions, and shaded areas the 90% moving block bootstrap bands, from the 0.05 to the 0.95 quantile of 1,000 bootstrap estimates (block length of 7).

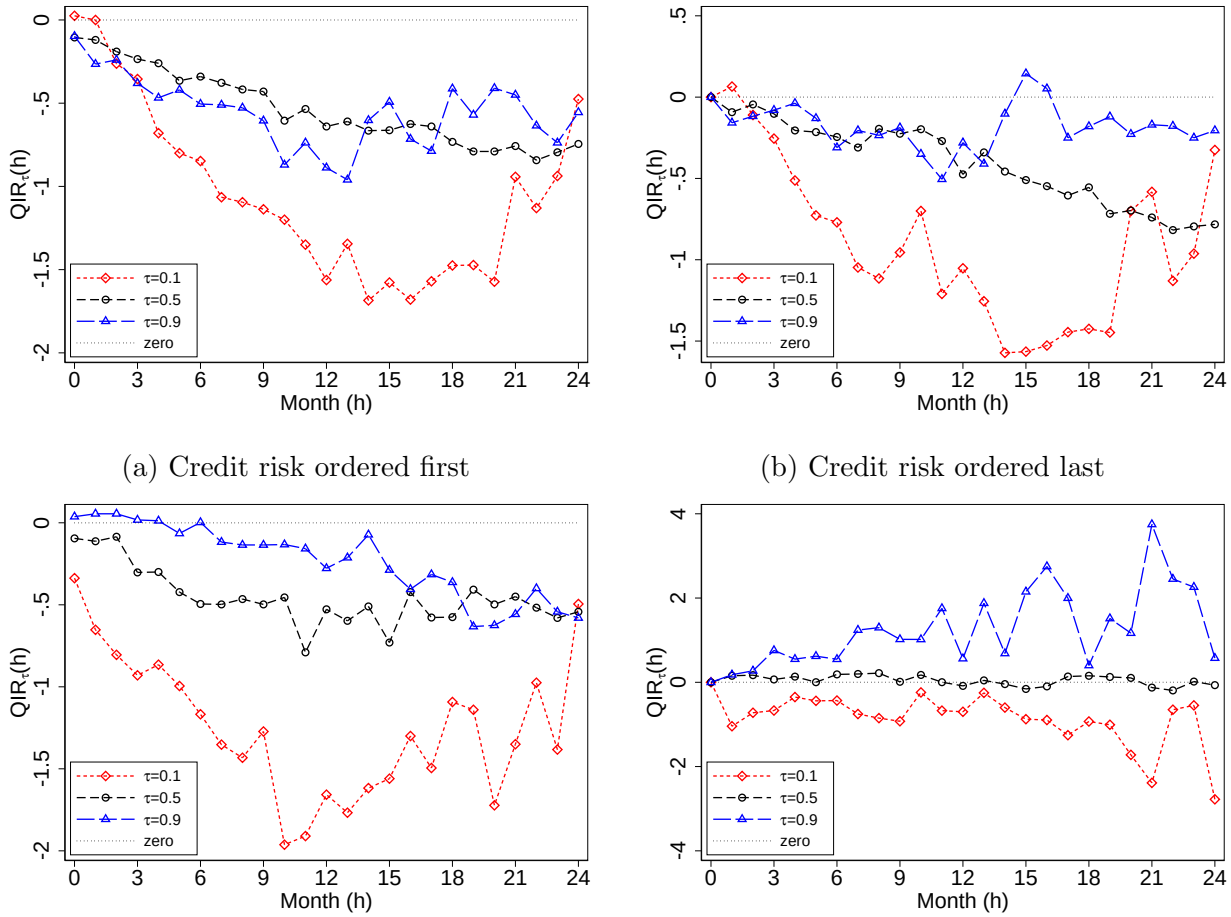
fits with a spline in the treatment at the six-month and one-year horizons, with the baseline controls.[21] For the lower decile and the median the fits are decreasing functions of the treatment. Between no move and a two-standard-deviation adverse move the lower decile

[21]The spline is the restricted cubic spline of appendix A2.3.3, with knots at -1.28, 0 and 1.28 standard deviations, and the two slopes are found by the same two-dimensional grid search. The level of each fit is the τ-quantile of the outcome net of the fitted treatment terms. Both slopes and the level are re-estimated in every bootstrap draw. As the nonlinear fits have heavy-tailed sampling distributions, the bands run from the 0.05 to the 0.95 quantile of the draws.

falls by between 2.0 and 3.8 percentage points across the four fits, 1.3 to 3.6 times as much as the median. Some point estimates bend, but in different directions. For example, the lower decile at six months falls by 3.8 points for an adverse move in volatility risk and rises by only 0.9 for a favorable one, while at one year it falls by 2.2 points for an adverse move in credit risk and rises by 3.4 for a favorable one. None of these bends is statistically significant: in all eight fits of the lower decile and the median, the 68% bootstrap interval of the coefficient on the second treatment term, which is zero when the structural quantile function is linear, contains zero.

The upper decile responds less to adverse moves than to favorable ones, except for credit risk at one year. In the other three fits, a two-standard-deviation adverse move changes it by at most 0.3 percentage points, while a favorable one raises it by between 1.0 and 4.0 points, the most for volatility risk at one year, from 5.7 to 9.7 percentage points. For volatility risk the 68% interval of the coefficient on the second treatment term excludes zero at both horizons, but none of the 12 bends is significant at the 10% level. In other words, the nonlinear fits confirm that the lower tail responds more than the median on both sides of zero, but the data cannot distinguish a bend of the lower tail like the one in the DSGE model from a linear response.

A3.3.2 Alternative timing restrictions



(a) Credit risk ordered first

(b) Credit risk ordered last

(c) Volatility risk ordered first

(d) Volatility risk ordered last

Figure 20: Cumulative response of Industrial Production (in % pts.) from a shock that increases credit (top panels) or volatility (bottom panels) risk by one standard deviation, plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the quantile impulse response. Note that the impact response (horizon $h = 0$) is by assumption zero when the shock is ordered after the dependent variable.

Ordering the treatment first changes the estimates little. Ordering it last leaves the credit risk responses similar, with a smaller lower-tail response. For volatility risk ordered last, however, the median response is close to zero and the upper-decile response becomes positive and grows with the horizon, which is implausible: the upper-decile response reaches 3.7 percentage points. With this ordering the controls include the stock market return of the

same month, and realized volatility rises with large returns of either sign, a nonlinearity that a first stage linear in the return cannot capture (lemma 2). Adding the absolute value and the square of the return to the first stage removes most of the effect: averaged over the 24 horizons, the upper-decile response falls from 1.3 to 0.3 percentage points.

A3.3.3 Spline controls

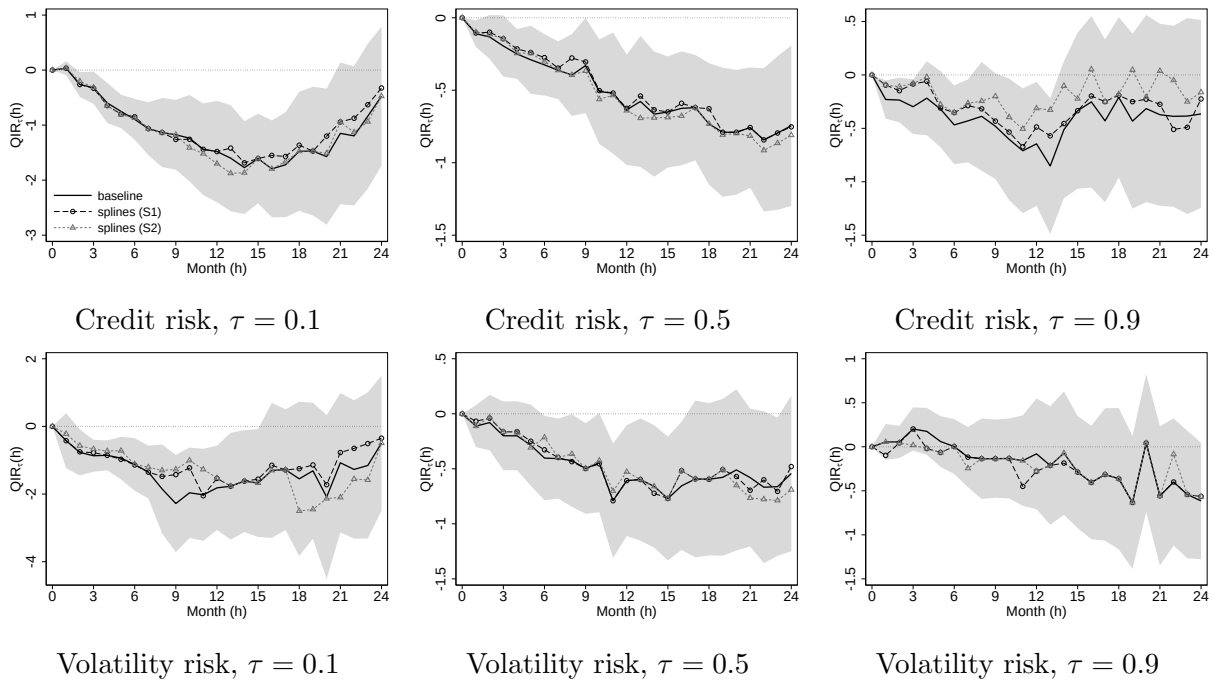


Figure 21: GQLP responses of industrial production (in % pts.) to a one-standard-deviation increase in credit risk (top row) and volatility risk (bottom row) with spline controls. The solid line and the gray area are the baseline estimate and its 90% moving block bootstrap band. Circles (S1): the current and lagged industrial production growth and the lagged shock enter GQLP's first stage as restricted cubic splines with five knots, the other controls linearly. Triangles (S2): every control enters as a restricted cubic spline with three knots.

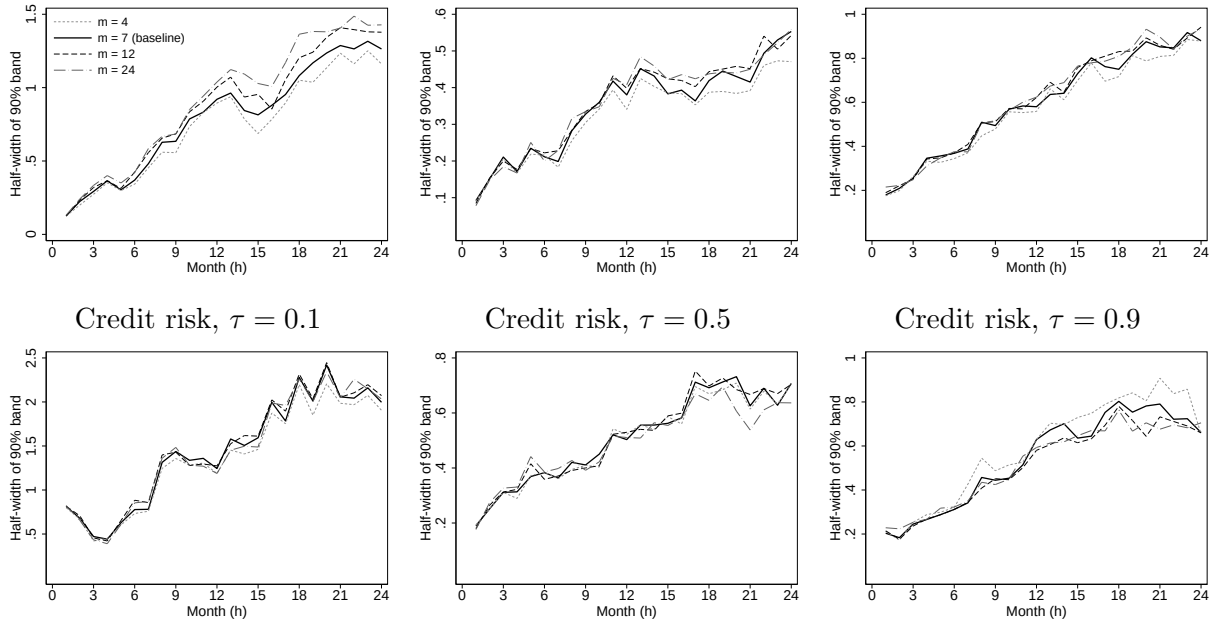
In the Monte Carlo designs the controls enter GQLP's first stage as restricted cubic splines, which lets the probability of the rank event depend nonlinearly on the state. Figure 21 repeats the empirical estimation with two such specifications, keeping the treatment, the outcome and the list of controls unchanged. In the first (S1), current and lagged industrial production growth and the first lag of the shock, the state variables most likely to matter for the lower tail, enter as restricted cubic splines with five knots and the other controls enter linearly. In the second (S2), every control enters as a restricted cubic spline with three knots. The figure compares their point estimates with the baseline estimates and 90% bands.

The point estimates stay inside the baseline 90% bands at every horizon and quantile, and the asymmetry is preserved. Over the first year, the response of the lower tail is 2.9 (S1)

and 2.7 (S2) times the median response for credit risk shocks, against 2.5 in the baseline, and 3.3 and 3.0 times for volatility risk shocks, against 3.7. The peak response of the $\tau = 0.1$ quantile to credit risk is 1.7 (S1) and 1.9 (S2) percentage points, against 1.8 in the baseline.

A3.3.4 Bootstrap block length

The baseline bands use blocks of $M = 7$ months. Because the h-step outcomes overlap, the dependence in the GQLP moment condition grows with the horizon, so longer blocks may be needed at long horizons. Figure 22 reports the half-width of the 90% bands for $M = 4, 12$ and 24. Averaged over horizons, the half-width changes by at most 18% for any shock, quantile and block length. The largest change is an 18% widening of the lower-tail bands of the credit risk responses with blocks of 24 months at horizons beyond a year; up to a year the changes lie between -6% and $+11\%$. None of the conclusions in section 4 depends on the block length.



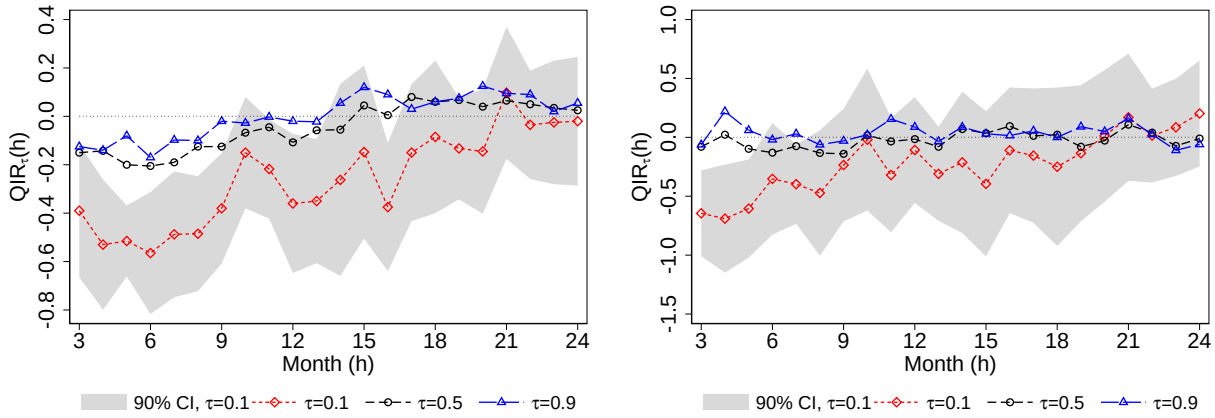
Credit risk, $\tau = 0.1$ | Credit risk, $\tau = 0.5$ | Credit risk, $\tau = 0.9$

Volatility risk, $\tau = 0.1$ | Volatility risk, $\tau = 0.5$ | Volatility risk, $\tau = 0.9$

Figure 22: Half-width of the 90% moving block bootstrap band of the GQLP response of industrial production to a one-standard-deviation increase in credit risk (top row) and volatility risk (bottom row), for block lengths $M = 4, 7, 12, 24$ months. The baseline uses $M = 7$. $B = 1{,}000$ bootstrap replications for every block length.

A3.3.5 Non-cumulative outcome

Figure 23 reports the GQLP responses of three-month industrial production growth ending at $t+h$, $100\,[\log(IP_{t+h}) - \log(IP_{t+h-3})]$, a non-cumulative outcome, with the baseline treatments and controls. These responses do not add up to the cumulative ones, since quantiles of sums are not sums of quantiles, but they show whether the asymmetry is a by-product of cumulating. It is not. Averaged over $h = 3, \ldots, 12$, the response of the $\tau = 0.1$ quantile of three-month growth is 3.0 times the median response for credit risk and 5.8 times for volatility risk, against 1.6 for both shocks with QLP. The lower-tail responses are largest about half a year after the shock (-0.57 percentage points at $h = 6$ for credit risk and -0.69 at $h = 4$ for volatility risk), and their 90% bands exclude zero at most horizons of the first year for credit risk and at four of them for volatility risk.

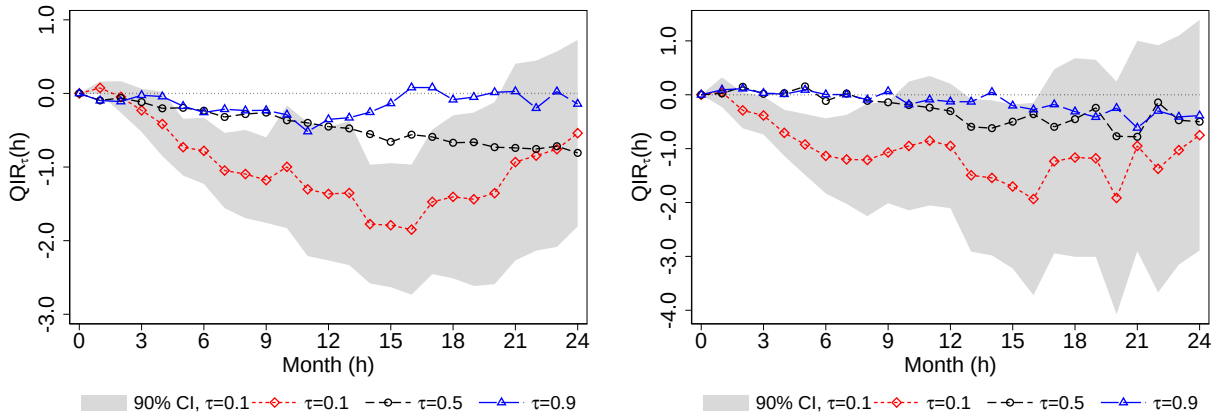


(a) Credit risk, three-month growth

(b) Volatility risk, three-month growth

Figure 23: Non-cumulative outcome. GQLP responses of three-month industrial production growth ending at $t + h$ (in % pts.) to a one-standard-deviation increase in credit risk (a) and volatility risk (b), $h = 3, \ldots, 24$, with the 90% moving block bootstrap band for $\tau = 0.1$. Quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$.

A3.3.6 Excluding the pandemic



(a) Credit risk, April 1973 to December 2019 (b) Volatility risk, April 1986 to December 2019

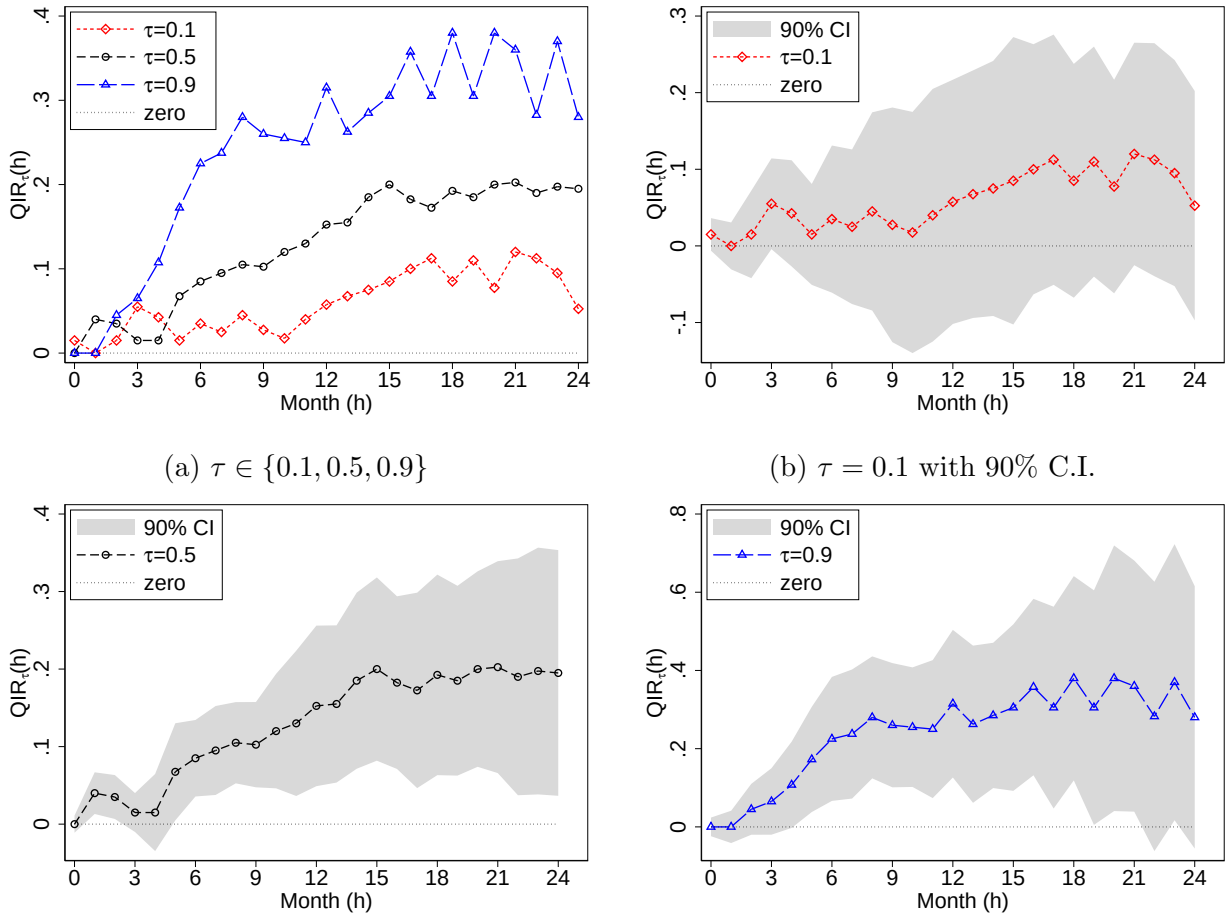
Figure 24: GQLP responses of industrial production (in % pts.) to a one-standard-deviation increase in credit risk and volatility risk with samples that end in December 2019, and the 90% moving block bootstrap band for $\tau = 0.1$.

The baseline samples include the pandemic, when industrial production fell by 17 percent between February and April 2020. Figure 24 repeats the estimation with samples that end in December 2019, keeping the baseline treatments, controls and standardization, and dropping the outcomes that reach into 2020. For credit risk the responses are close to the baseline. Over the first year the response of the $\tau = 0.1$ quantile averages -0.76 percentage points, against -0.86 in the full sample, and it is 3.1 times the median response, against 2.5. Its trough, -1.85 percentage points after 16 months, is almost the same as in the baseline, and its 90% band excludes zero from the fifth to the twentieth month. For volatility risk the lower-tail response builds up more slowly: it averages -0.80 percentage points over the first year, against -1.36, and reaches its trough of -1.93 after 16 months, against -2.28 after 9 months in the baseline. Its band excludes zero at most horizons between the third and the sixteenth month. The median and upper-decile responses, however, stay close to zero over the first year, so the asymmetry remains. In other words, the pandemic makes the lower-tail response to volatility risk larger and faster, but it is not the source of the asymmetry.

A3.3.7 Unemployment

Figures 25 and 27 replace industrial production with the unemployment rate, and Figures 26 and 28 report its responses for quantiles from $\tau = 0.05$ to $\tau = 0.95$ at four horizons, as Figures 15 and 16 do for industrial production. The outcome is the change in the unemployment rate between $t - 1$ and $t + h$, in percentage points, and the treatments and controls are those of the baseline. Since the unemployment rate is not among the controls, its response on impact is not restricted to zero. A rise in unemployment is the bad outcome, so the tail of interest is now the upper one.

The results mirror those for industrial production. Over the first year the response of the $\tau = 0.9$ quantile averages 0.18 percentage points for credit risk and 0.34 for volatility risk, 2.3 and 5.1 times the median response, and it peaks at 0.38 and 0.55 percentage points after 18 and 19 months. Its 90% band excludes zero from the fifth to the twenty-first month for credit risk and from the fourth to the sixteenth month for volatility risk, while the $\tau = 0.1$ quantile barely moves. Across quantiles the largest response is above the median at every horizon, at a quantile between 0.70 and 0.95. For example, after 18 months it is 0.13 at $\tau = 0.05$, 0.19 at the median and 0.51 at $\tau = 0.95$ for credit risk, and 0.06, 0.19 and 1.43 for volatility risk, while the response of the mean estimated by local projections is 0.32 and 0.29. QLP finds almost no asymmetry: over the first year the response of the $\tau = 0.9$ quantile is 1.3 times the median response for credit risk and 1.2 times for volatility risk.



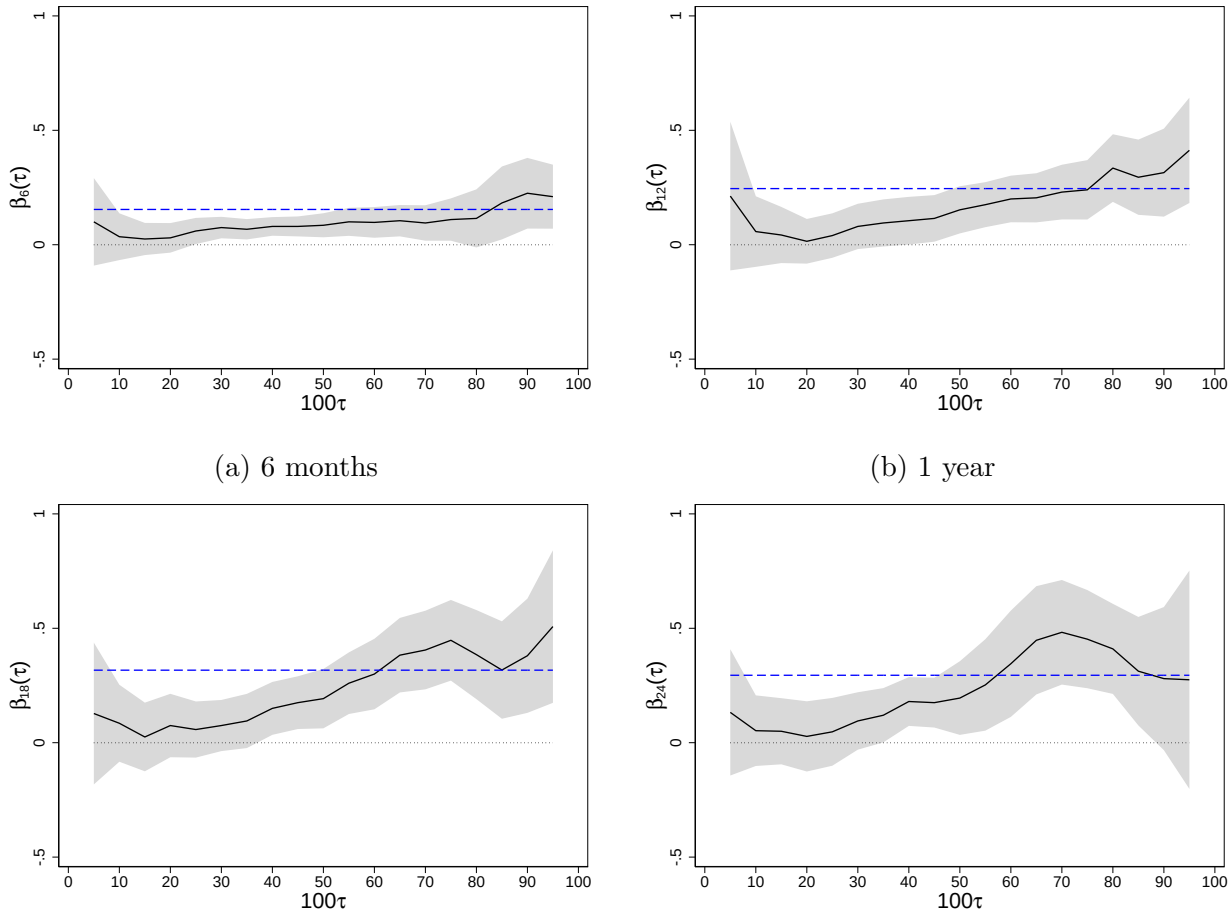
(a) $\tau \in \{0.1, 0.5, 0.9\}$

(b) $\tau = 0.1$ with 90% C.I.

(c) $\tau = 0.5$ with 90% C.I.

(d) $\tau = 0.9$ with 90% C.I.

Figure 25: Cumulative response of the unemployment rate (in % pts.) from a shock that increases credit risk by one standard deviation, plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the quantile impulse response. Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). The impact response (horizon $h = 0$) is not restricted, since the unemployment rate is not among the controls.



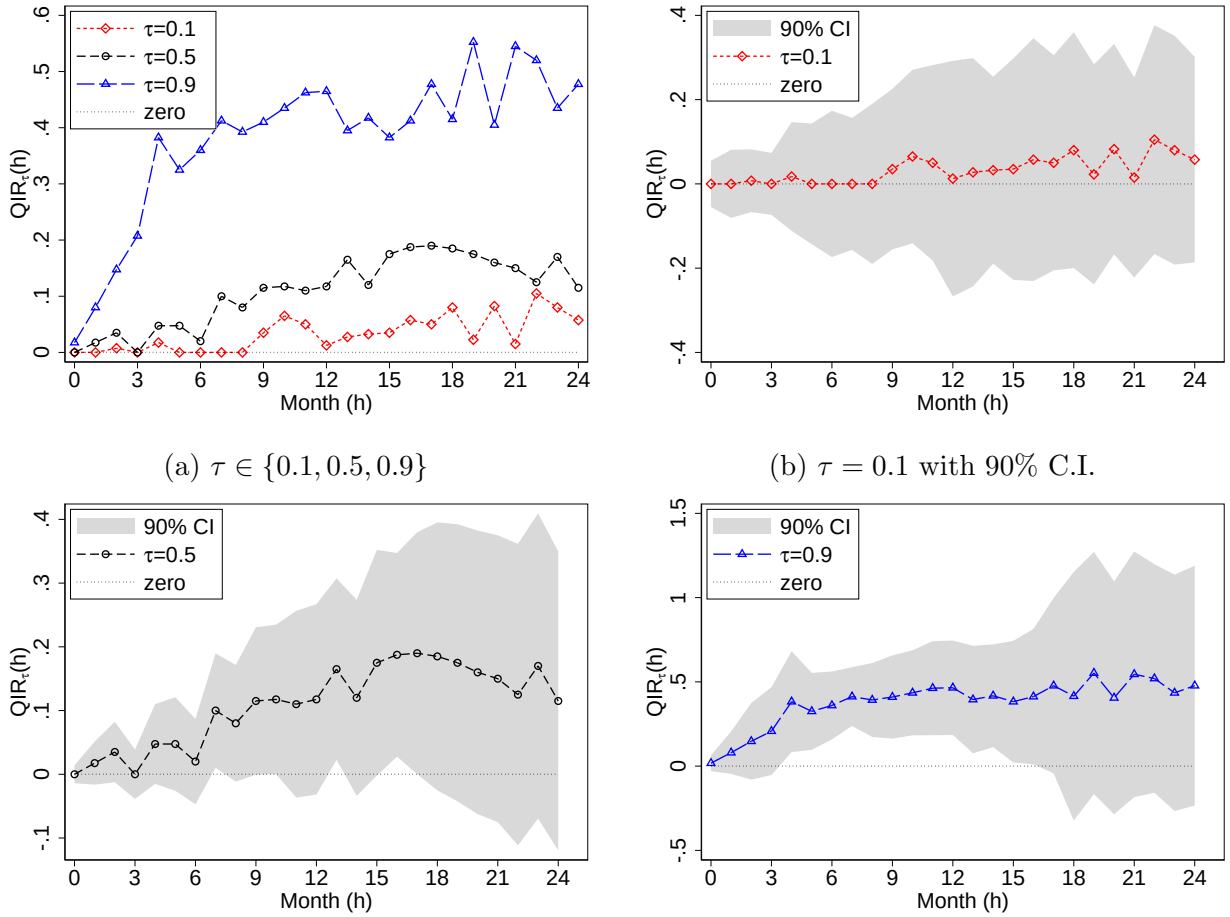
(a) 6 months

(b) 1 year

(c) 18 months

(d) 2 years

Figure 26: Responses of the unemployment rate (in % pts.) to a shock that increases credit risk by one standard deviation, plotted for four horizons $h \in \{6, 12, 18, 24\}$ (panels from top left to bottom right). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Blue dashed line reports the response of the mean estimated from conventional local projections.



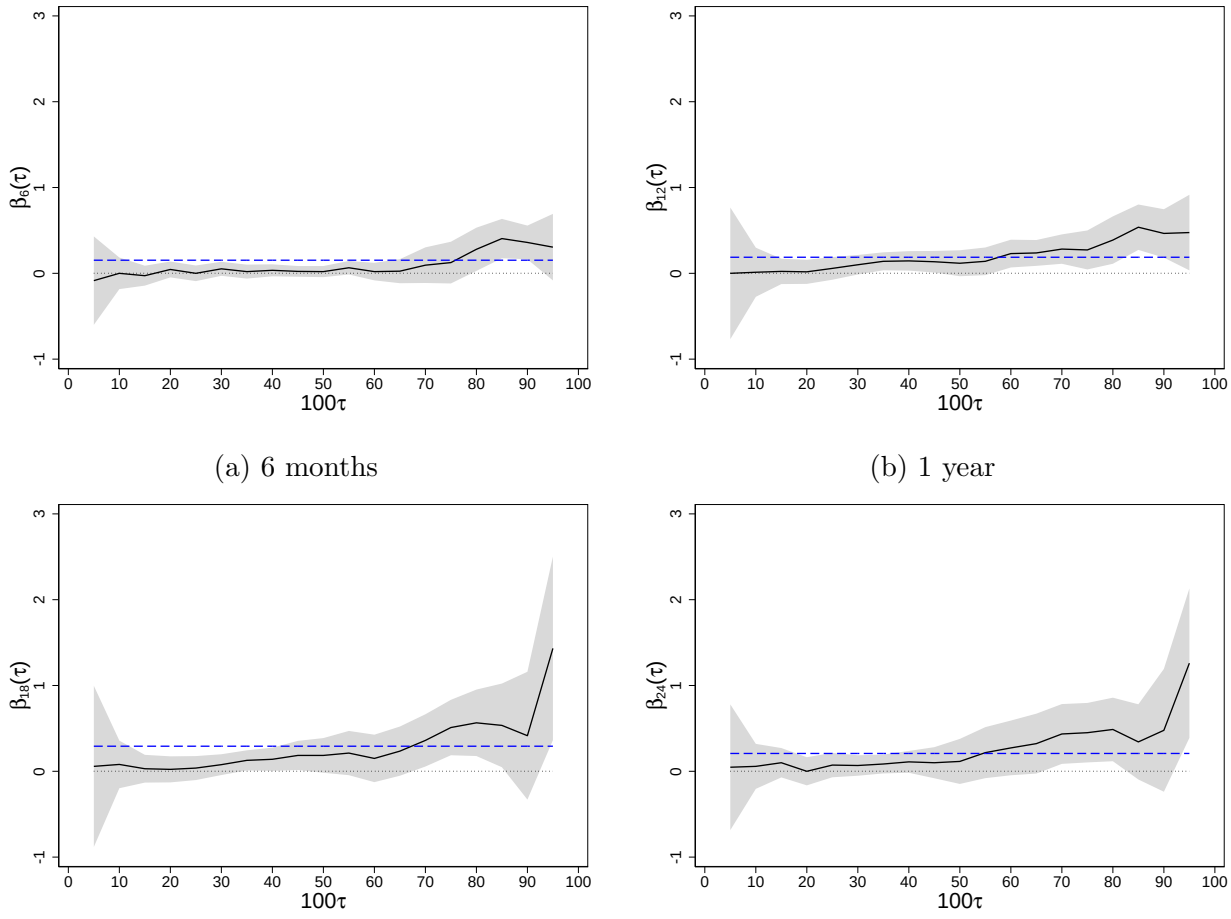
(a) $\tau \in \{0.1, 0.5, 0.9\}$

(b) $\tau = 0.1$ with 90% C.I.

(c) $\tau = 0.5$ with 90% C.I.

(d) $\tau = 0.9$ with 90% C.I.

Figure 27: Cumulative response of the unemployment rate (in % pts.) from a shock that increases volatility risk by one standard deviation, plotted for three quantiles $\tau \in \{0.1\diamond, 0.5\circ, 0.9\triangle\}$. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the horizon h in months. Dashed lines plot the quantile impulse response. Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). The impact response (horizon $h = 0$) is not restricted, since the unemployment rate is not among the controls.



(a) 6 months

(b) 1 year

(c) 18 months

(d) 2 years

Figure 28: Responses of the unemployment rate (in % pts.) to a shock that increases volatility risk by one standard deviation, plotted for four horizons $h \in \{6, 12, 18, 24\}$ (panels from top left to bottom right). The responses were estimated for quantiles from $\tau = 0.05$ to $\tau = 0.95$ in 0.05 increments. Y-axis is the estimated response $\hat{\beta}_h(\tau)$, x-axis is the quantile τ (multiplied by 100 for legibility). Shaded area is the moving block bootstrap 90% Confidence Interval (with block length of 7, and 1,000 bootstrap replications). Blue dashed line reports the response of the mean estimated from conventional local projections.