Coherent shocks: External identification and internal validation with an application to geopolitical risk*

Yves Schüler
Deutsche Bundesbank

Sarah Arndt
European Central Bank

Yevheniia Bondarenko
Deutsche Bundesbank

Vivien Lewis
Deutsche Bundesbank & CEPR

Matthias Rottner
BIS & Deutsche Bundesbank

July 29, 2026

Abstract

We propose the concept of coherent shocks: externally identified shocks whose structural interpretation is independently validated within the empirical model. Using textual data, we classify articles by economic mechanism and construct mechanism-specific shock proxies. Our coherence criterion measures the posterior share of impulse-response draws matching the assigned signature, benchmarked against its prior share. Applied to geopolitical risk (GPR), we use a large language model to construct supply-type and demand-type GPR series. Shocks identified from both series are strongly coherent with their assigned signatures. Innovations in the overall GPR index blend the two, and neither recursive nor sign-restriction identification separates them. The criterion applies to any externally identified shock for which theory implies a response signature.

Keywords: coherent shocks, shock validation, external identification, Bayesian VAR, large language models, geopolitical risk

JEL classification: C11, C32, E31, E32, F51.

*We thank Christiane Baumeister, Leland Bybee, Anton Korinek, Yasin Mimir, Alex Popov, Harald Uhlig and Chien-Wen Yang. We also thank seminar participants at Deutsche Bundesbank, the 2025 AWG-MPAG Annual Workshop, the 2026 meeting of the VfS Monetary Policy Committee, the 2026 IAAE Annual Conference, and the 101st WEAI Annual Conference for helpful comments. Simone Ferretti provided outstanding research assistance. The views expressed in this paper are those of the authors and do not necessarily coincide with the views of the Deutsche Bundesbank, the European Central Bank, the Bank for International Settlements (BIS), or the Eurosystem.

1 Introduction

Over more than forty years, identifying structural shocks has been a central project of empirical macroeconomics, using either restrictions imposed within the empirical model (Sims, 1980; Blanchard and Quah, 1989) or information brought from outside the model (Romer and Romer, 1989; Kuttner, 2001; Stock and Watson, 2012).[1] While identification has received much attention, validation has received far less. Once a shock is identified, its interpretation is often inherited from the identifying strategy itself, rather than independently evaluated. This emphasis on identification over validation is partly structural. Within-model identification imposes theoretical restrictions to recover the shock, so those same restrictions cannot subsequently serve as an independent check. External identification avoids this circularity by using information from outside the model, but the shock's structural interpretation typically derives from the external information and the assumptions used to isolate it.

This paper combines identification with validation: we identify shocks outside the model and validate their structural interpretation inside it. Outside the model, we use textual data to classify events by the economic mechanism they describe, the process through which an event is reported to affect the economy, and map each mechanism to a structural shock type, yielding mechanism-specific shock proxies. Inside the model, we evaluate whether the impulse responses to these proxies are consistent with the theoretical response pattern associated with each assigned shock type, which we refer to as its signature. Because these signatures are used neither to construct the proxies nor to recover the shocks, they remain available as independent, testable implications. We formalize this validation through a coherence criterion that measures the posterior share of impulse-response draws consistent with the assigned signature, benchmarked against the corresponding prior share. The more strongly the data favor impulse responses that match the assigned signature, the more coherent the shock is.

We illustrate our approach using geopolitical risk (GPR), building on the seminal news-based index of Caldara and Iacoviello (2022), whose innovations are widely used as proxies for geopolitical risk shocks in empirical work. GPR is a natural laboratory for our framework because the GPR index is topic-defined rather than shock-defined: it captures the intensity of newspaper coverage of geopolitical risk, but does not distinguish among the structural shock types embedded in that coverage. Adverse geopolitical events can operate through supply-side mechanisms, by raising production and input costs, disrupting supply chains, or reducing the availability of goods; they can also operate through

[1] Within-model approaches include recursive timing restrictions (Sims, 1980), long-run restrictions (Blanchard and Quah, 1989), and sign restrictions (Uhlig, 2005; Rubio-Ramírez et al., 2010). External approaches include narrative identification (Romer and Romer, 1989, 2004; Ramey, 2011), proxy methods (Stock and Watson, 2012; Mertens and Ravn, 2013; Stock and Watson, 2018; Plagborg-Møller and Wolf, 2021), and high-frequency identification (Kuttner, 2001; Gürkaynak et al., 2005; Gertler and Karadi, 2015).

demand-side mechanisms, by eroding confidence and curtailing spending. Economic theory associates the two shock types with distinct response signatures: after an adverse supply shock, output should fall while prices rise; after an adverse demand shock, both should fall.

Using the Caldara–Iacoviello GPR article corpus and search methodology as a starting point, we classify news articles according to the economic mechanism they describe and construct separate supply and demand GPR indices. Specifically, we employ a large language model to classify each article using a structured questionnaire that determines whether it describes a supply or demand mechanism; an article may describe both or neither. The resulting decomposition is economically intuitive: events such as 9/11 are classified primarily as demand-type episodes, while the onset of Russia's full-scale invasion of Ukraine is classified predominantly as supply-type. In a Bayesian VAR, innovations in the two indices generate distinct and theory-consistent responses, differing in the price response, the monetary policy reaction, and the timing of real effects. The coherence criterion quantifies this validation: the shocks recovered from the two proxies are coherent with their assigned types, while the data provide no support for the opposing interpretation.

Can within-model identification recover this distinction from the Caldara–Iacoviello GPR index alone? We find that it cannot. Under a recursive ordering, the structural interpretation of the shock identified from the GPR index depends on the price measure used: evaluated against core PCE, the shock is coherent with a demand signature; evaluated against broad PCE, it is coherent with a supply signature. This instability is consistent with a shock that mixes the two. Sign restrictions do not separate the shocks either. The two sign-restricted shocks differ only in the responses whose signs are imposed; otherwise, their impulse responses are indistinguishable. Since the demand and supply signs are imposed for identification, they cannot also validate the resulting interpretation; accordingly, the coherence criterion registers no evidence beyond the prior for either interpretation. Innovations in the topic-defined GPR index therefore mix different structural shocks, illustrating the limits of what the macroeconomic covariance structure alone can recover once articles describing different economic mechanisms have been combined in a single news index.

The paper contributes to three literatures. First, we contribute to the literature on structural shock identification, and in particular to external and narrative identification (Romer and Romer, 1989, 2004; Ramey, 2011; Stock and Watson, 2012; Mertens and Ravn, 2013). We extend this tradition by classifying events according to the economic mechanism they describe, rather than by their exogeneity or episode type.[2] Our central

[2]Narrative information can also be used inside the model: Antolín-Díaz and Rubio-Ramírez (2018) impose restrictions on the sign of a shock, or on its contribution to a historical episode, as additional identifying restrictions. In our approach the textual information enters outside the model only.

contribution, however, shifts the focus from identification to validation of the structural interpretation assigned to an identified shock. Our approach uses textual data to construct shock proxies from mechanism-based event classifications. Because this classification takes place before any macroeconomic model is estimated, the theoretical response signature of the assigned shock type is not imposed for identification and remains available as an implication that can be evaluated in the model. We formalize this validation step through a coherence criterion, defined as the posterior share of impulse responses consistent with the assigned shock type, benchmarked against the prior.

Second, the paper contributes to the literature using news-based indices in macroeconomics and finance (e.g., Baker et al., 2016; Caldara and Iacoviello, 2022; Engle et al., 2020). Such indices are typically topic-defined: they measure the intensity of news coverage about uncertainty, geopolitical risk, climate risk, or related concepts. Their innovations need not correspond to a single structural shock. We show how classifying the underlying text by economic mechanism can be used to construct mechanism-specific shock proxies and provide explicit control over the disturbance each proxy is intended to measure. More broadly, our classification step relates to a growing literature using textual methods, including machine learning and large language models, to construct or refine empirical shock measures or recover economically interpretable components from text (e.g., Acosta, 2026; Aruoba and Drechsel, 2026; Doh et al., 2026; Handlan, 2022; Kwon et al., 2025; Lumbanraja et al., 2026). Relative to this literature, our contribution lies in what follows the textual classification: we construct the shock proxy outside the model and reserve its theoretical response signature for an independent test inside it. The result is a news-based shock proxy whose structural interpretation is not assumed but tested through our coherence criterion.

Third, the paper contributes to the literature on geopolitical risk, which is surveyed in Mohr and Trebesch (2025). The seminal GPR index of Caldara and Iacoviello (2022) made geopolitical risk measurable by tracking geopolitical threats and acts in the news. Building on this contribution, we show that the events captured by the GPR index can be decomposed into demand- and supply-type geopolitical risk components. This decomposition matters because the two types of geopolitical events transmit differently to output, prices, and monetary policy. It also provides real-time indicators of demand- and supply-type geopolitical risk, which are useful for macroeconomic analysis and policy assessment. We also contribute to a growing literature proposing refinements to the measurement of geopolitical risk (e.g., Hassan et al., 2019; Bondarenko et al., 2024; Clayton et al., 2025; Iacoviello and Tong, 2026).

2 External identification and internal validation: A framework

This section develops our two-step framework. First, we identify shocks outside the macroeconomic model, using textual information to classify events by the economic mechanism they describe. This step can in principle be applied to any text corpus in which events are described together with their economic mechanisms. In our application, the corpus consists of newspaper articles on geopolitical risk, selected through a keyword search. Second, once the resulting shock proxies are brought into a macroeconomic model, the model is used for validation rather than identification. Because the signature was not imposed to recover the shock, it remains available as a testable implication.

2.1 External identification through mechanism-based text classification

The external-identification step uses textual information to construct shock proxies before any macroeconomic model is estimated. The key premise is that text often describes not only an event itself, but also the economic mechanism through which it affects the economy. These mechanisms provide the identifying information: an event described as disrupting production, raising input costs, or reducing the availability of goods points to a different structural shock type than an event described as reducing confidence or spending.

We now apply this idea to news-based indices, the setting of our empirical application. Most news-based macroeconomic indices are constructed by applying keyword searches to large newspaper corpora. Let $\mathcal{U}_t$ denote the set of all articles in the underlying corpus in period t, and let $\mathcal{A}_t \subseteq \mathcal{U}_t$ denote the subset selected by a given search query. A typical newspaper-based index is the share of selected articles,

$$X_t = \frac{1}{|\mathcal{U}_t|} \sum_{a_{i,t} \in \mathcal{U}_t} \mathbf{1}\{a_{i,t} \in \mathcal{A}_t\}, \tag{1}$$

where $|\mathcal{U}_t|$ is the total number of articles in period t and the indicator equals one if article $a_{i,t}$ satisfies the search query; equivalently $X_t = |\mathcal{A}_t|/|\mathcal{U}_t|$. Such indices are widely used as empirical proxies for latent economic conditions, including economic policy uncertainty, geopolitical risk, and climate-related risk (Baker et al., 2016; Caldara and Iacoviello, 2022; Engle et al., 2020).[3] In macro-financial applications, it is the *innovation* in such an index, its unexpected component, that serves as the empirical shock proxy.

In this setting, the relevant measurement problem is that a keyword-based index

[3] A broad collection of newspaper-based indicators covering a diverse set of topics is available at https://www.policyuncertainty.com/ (last accessed: June 9, 2026).

is topic-defined rather than shock-defined. A single search query can therefore select articles that describe different economic mechanisms, alongside articles that describe no mechanism at all. Formally, the innovation in the index,

$$\nu_t^X = X_t - \mathbb{E}\left[X_t \mid \mathcal{I}_{t-1}\right], \tag{2}$$

where $\mathcal{I}_{t-1}$ is the information set available prior to period t, generally blends the contributions of several structural shocks with non-structural news content.

An empirical model that treats ν_t^X as the innovation to a single shock will then recover a combination of the underlying disturbances, weighted toward the shock types that appear most prominently in the corpus, and cannot remove the non-structural component. Once articles describing different economic mechanisms have been combined in a single news index, conventional within-model identification need not recover the underlying structural distinction.[4] Rather than ask an empirical model to undo this conflation, we address it at the measurement stage: the mechanism classification developed below assigns articles to shock types before any model is estimated, so each resulting index is associated with a single shock type.

Mechanisms as the identifying property. To separate the shocks embedded in the index, we exploit that each structural shock is described in the text through one or more observable economic mechanisms, i.e. economic processes that an article attributes to a specific event. A negative supply shock may be described through rising input costs or prices, or disruptions to the supply of goods; a negative demand shock through declining confidence or curtailed spending. We treat these mechanisms as the property that identifies the shock, in the same spirit in which recursive schemes use timing and sign-restriction schemes use response signs.

Let $m \in \{1, \dots, M\}$ index the mechanisms and let $\kappa(m) = k$ denote the mapping from mechanism m to the structural shock type k to which it points. The mapping is many-to-one: several mechanisms may point to the same shock type, but each mechanism points unambiguously to one shock type. This is the central design requirement of the classification. For each article $a_{i,t}$ and mechanism m, let $c_m(a_{i,t}) \in \{0, 1\}$ indicate whether the article is classified as describing mechanism m.

[4]Here, we refer to conventional within-model identification schemes, including recursive, long-run, and sign-restriction approaches. Identification schemes exploiting heteroskedasticity can statistically separate structural disturbances without conventional economic restrictions (see, e.g., Rigobon, 2003), but recover them unlabeled. Their economic interpretation, therefore, still requires information from outside the covariance structure; the coherence criterion below is one such source.

Constructing mechanism-specific indices. We assign an article to shock k whenever it describes at least one mechanism mapping to k, and construct the share

$$X_{k,t} = \frac{1}{|\mathcal{U}_t|} \sum_{a_{i,t} \in \mathcal{U}_t} \mathbf{1}\{a_{i,t} \in \mathcal{A}_t\} \, \mathbf{1}\Big\{\textstyle\sum_{m:\kappa(m)=k} c_m(a_{i,t}) \geq 1\Big\}. \tag{3}$$

Within a given shock, each article is counted at most once, regardless of how many of that shock's mechanisms it describes. Across shocks, an article may enter more than one index if it describes mechanisms mapping to different shocks; this reflects genuine events that transmit through more than one structural shock type. Articles that describe none of the classified mechanisms enter neither index and make up the residual share.

The object that proxies a structural shock is, again, the innovation in the corresponding index,

$$\nu_{k,t} = X_{k,t} - \mathbb{E}[X_{k,t} \mid \mathcal{I}_{t-1}]. \tag{4}$$

Unlike ν_t^X, this innovation is associated with a single shock type by construction. Because the mechanism-specific indices overlap, the innovation ν_t^X does not admit a simple additive decomposition into the $\nu_{k,t}$: a single article describing mechanisms of more than one shock contributes to several of them. The decomposition describes the shock processes embedded in the index, not a partition of the underlying articles.

The approach is subject to one fundamental limitation: a shock can be identified only insofar as its economic mechanism is observable in the textual source. Mechanisms that the source does not describe, or describes too ambiguously to be classified, fall outside what the method can recover; the choice of textual source therefore determines which shocks are identifiable. Whether $\nu_{k,t}$ can in turn serve as a proxy for the structural shock $\varepsilon_{k,t}$, and under what conditions, is the subject of the next subsection.

2.2 Internal validation through coherence

We now turn from external identification to internal validation. Under what conditions can an externally constructed shock proxy be interpreted as capturing the structural shock type it is assigned to, and how can this interpretation be evaluated inside the macroeconomic model? We first state these conditions for the mechanism-specific indices constructed above and then introduce a coherence criterion that applies to any externally identified shock.

From the mechanism-specific index to the structural shock. In our text-based setting, the mechanism classification isolates coverage associated with a particular structural shock type before any macroeconomic model is estimated. The resulting innovation can therefore be used in standard empirical frameworks. For concreteness, we consider a reduced-form VAR; the criterion itself only requires posterior draws of impulse responses, however obtained.[5] Collect the macroeconomic variables together with one mechanism-specific index $X_{k,t}$ in the vector Y_t, and let

$$Y_t = B(L)\, Y_{t-1} + u_t, \tag{5}$$

where u_t denotes the vector of reduced-form innovations. In this setting, the reduced-form innovation in the index equation is the empirical counterpart of the mechanism-specific innovation $\nu_{k,t}$ defined above.

Let the reduced-form and structural innovations be related through

$$u_t = A\varepsilon_t, \tag{6}$$

where A is the contemporaneous impact matrix. To recover the dynamic effects associated with $\varepsilon_{k,t}$, we order $X_{k,t}$ first in a recursive VAR. This implementation is closely related to the internal-instrument representation of Plagborg-Møller and Wolf (2021), which makes explicit the conditions for a structural interpretation of the index innovation.

The relevant conditions are analogous to the standard relevance and exclusion conditions in the external-instrument literature (Stock and Watson, 2012; Mertens and Ravn, 2013; Stock and Watson, 2018). First, the innovation $\nu_{k,t}$ must contain information about the target structural shock $\varepsilon_{k,t}$. Second, it must be contemporaneously orthogonal to other structural shocks. Under these conditions, ordering $X_{k,t}$ first identifies the impulse responses associated with $\varepsilon_{k,t}$, normalized to a one-standard-deviation innovation in the index.

The construction of the index is designed to make these conditions plausible. The initial keyword search selects contemporaneous news coverage of the underlying events, while the classification retains only articles that describe a mechanism of the target shock type. This strengthens the relevance condition: rather than relying on an incidental correlation, the index is constructed to track the target disturbance directly. The maintained timing assumption is that the resulting index innovation reflects the underlying event rather than contemporaneous macroeconomic disturbances. The maintained exclusion assumption is that the classification sufficiently limits contamination from other shock types and from non-structural news content. Finally, ordering the index innovation first in a recursive VAR follows common practice in this literature (see, among others, Baker et al., 2016; Caldara and Iacoviello, 2022; Bondarenko et al., 2024). In our setting, this ordering is supported by an additional layer of identifying information: the mechanism classification isolates news coverage associated with a particular shock type before estimation.

[5] This requirement is what currently excludes standard local projections: they deliver pointwise estimates equation by equation rather than joint posterior draws across variables and horizons, so the share of draws satisfying a multivariate signature is not defined. Extending the criterion to that setting requires joint inference across response equations, for instance, through system estimation or a joint bootstrap. We leave this to future work.

These assumptions cannot be verified from the textual classification alone. The coherence criterion introduced below therefore provides an internal validation exercise: it evaluates whether the dynamic responses implied by the externally constructed proxy are consistent with the signature of the structural shock type it is intended to capture.

Posterior coherence share. The central idea is that an identified shock should behave in line with the signature of its assigned type. Let $\mathcal{C}$ denote a set of theoretical restrictions on the impulse responses across variables and for specific horizons. In general, these may involve sign, relative-magnitude, persistence, or timing restrictions implied by the shock type. In our application, $\mathcal{C}$ consists of sign restrictions: output and prices declining together under a demand shock, and moving in opposite directions under a supply shock.

For posterior draw d, define the coherence indicator

$$\chi_d = \begin{cases} 1 & \text{if draw } d \text{ satisfies } \mathcal{C}, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

and the posterior coherence share

$$C = \frac{1}{D} \sum_{d=1}^{D} \chi_d, \tag{8}$$

where D is the number of posterior draws. The posterior coherence share measures the extent to which the identified shock aligns with the signature of its assigned type.

Coherence criterion. The posterior share C is informative only relative to what the model would produce before the parameters are updated by the likelihood. We therefore compute the prior-benchmark coherence C_0, the share of impulse responses drawn from the calibrated prior distribution of the VAR parameters that satisfy $\mathcal{C}$. Comparing the two yields a Bayes-factor-style diagnostic,

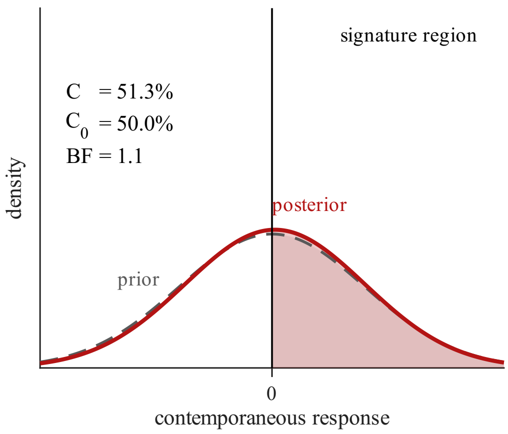
$$\text{BF} = \frac{C/(1-C)}{C_0/(1-C_0)}, \tag{9}$$

the factor by which likelihood updating shifts the odds in favor of theoretical consistency relative to the prior benchmark. Values below one indicate that the data shift probability mass away from the theoretically coherent region. As an interpretive reference scale, we adopt the conventional Bayes-factor categories of Kass and Raftery (1995): values between 3 and 20 constitute positive evidence, between 20 and 150 strong evidence, and above 150 very strong evidence.[6] We describe a shock accordingly as positively, strongly,

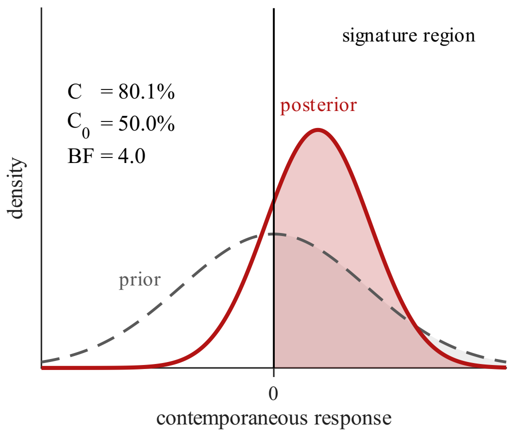
[6] Because our prior hyperparameters are calibrated on the data, following the Minnesota-prior practice in the Bayesian VAR literature, C_0 is a prior-benchmark diagnostic rather than a fully data-free reference.

or very strongly coherent with its assigned type as the BF exceeds 3, 20, or 150; below 3, we take the data to provide no evidence for the assigned interpretation.

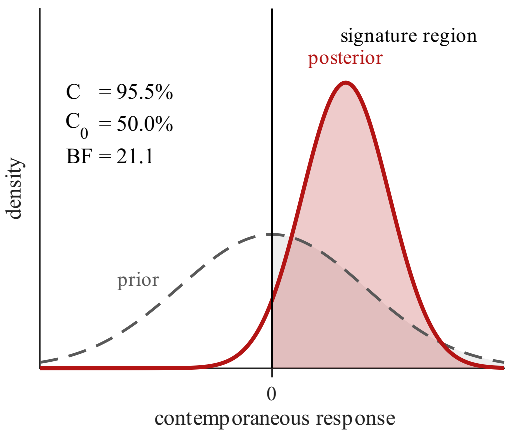
This comparison has content only because the macroeconomic responses are left unrestricted. A shock identified by imposing the signs in $\mathcal{C}$ would satisfy them by construction, driving C toward one and the prior benchmark equally high, so that BF stays near one regardless of the data. Here C is free to be low, and a posterior coherence well above its prior benchmark is genuine evidence that the classification has isolated a shock that behaves as its interpretation requires.



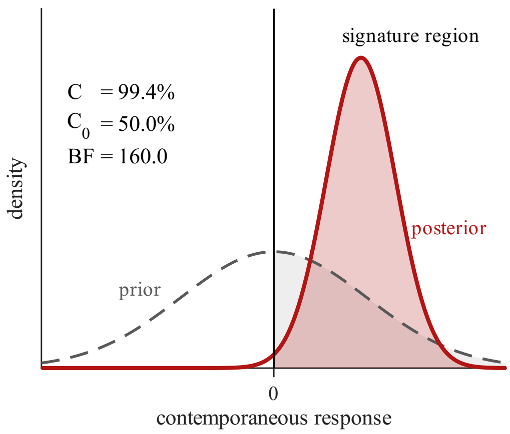
(a) No evidence (BF = 1.1)



(b) Positively coherent (BF = 4.0)



(c) Strongly coherent (BF = 21.1)



(d) Very strongly coherent (BF = 160.0)

Figure 1: The coherence criterion: prior, posterior, and the signature region

Notes: Each panel shows a schematic prior (grey, dashed) and posterior (red, solid) density for the contemporaneous response of a single variable. The densities are illustrative, not estimated. The panels restrict a single response. In this example, the signature region is the half-line to the right of zero. The prior share C_0 is the grey area in the region, the posterior coherence C the red area, and BF $= [C/(1-C)]/[C_0/(1-C_0)]$ the factor by which likelihood updating raises the odds of the region relative to the prior. Across the panels the prior is held fixed, so $C_0 = 50\%$ throughout, and only the posterior moves.

We treat the resulting statistic accordingly, as a measure of how far likelihood updating moves coherence beyond the calibrated prior.

Figure 1 illustrates the criterion. Each panel plots a schematic prior density (grey, dashed) and posterior density (red, solid) for the contemporaneous response of one variable. In this example, the signature region is the set of responses of the sign the assigned shock type implies, here the half-line to the right of zero. The prior share C_0 is the grey mass in this region and the posterior coherence C the red mass; their odds ratio is the BF diagnostic. When the data are uninformative about the sign, the posterior basically coincides with the prior and the two masses are equal, so BF = 1.1 and the criterion reports no evidence for the assigned type (panel (a)). As the likelihood pulls the posterior into the signature region, its mass there exceeds the prior benchmark and BF rises: the shock is positively coherent at BF = 4.0 (panel (b)), strongly coherent at BF = 21.1 (panel (c)), and very strongly coherent at BF = 160.0.

Given an uninformative prior for the impulse responses, the benchmark C_0 depends on how demanding the signature is: restricting a single response leaves roughly half the prior mass in the region, but a signature that restricts several responses jointly has a much lower prior share, for instance, about 25% under two sign restrictions. A lower C_0 raises the Bayes factor for the same posterior coherence share, so signatures that restrict more responses are the more demanding tests.

Finally, the coherence criterion is not specific to news-based indices. It only requires an externally identified shock proxy, an assigned structural interpretation, and a theoretical response signature against which that interpretation can be evaluated.

3 From geopolitical risk to coherent supply and demand shocks

This section applies our general framework to geopolitical risk, building on the GPR index of Caldara and Iacoviello (2022). From their article corpus, we construct two mechanism-specific indices, a supply GPR index and a demand GPR index, by classifying each article according to the economic mechanism it describes, and study their macroeconomic effects. We then validate the structural interpretation of the resulting shocks with our coherence criterion, and contrast them with shocks identified from the Caldara–Iacoviello index itself, using both a recursive ordering and sign restrictions designed to separate supply from demand.

3.1 Constructing supply and demand GPR indices

We retrieve the newspaper articles using the search methodology underlying the geopolitical risk (GPR) indicator of Caldara and Iacoviello (2022). Their index is constructed as the share of newspaper articles matching a predefined search query covering geopolitical threats and adverse geopolitical events. We obtain the relevant articles from the

Factiva Analytics database for the period from January 1985 to March 2026. Our sample contains 233,597 articles from seven news outlets: *USA Today, Chicago Tribune, The Globe and Mail, The Daily Telegraph, The Guardian, The Wall Street Journal,* and *The Washington Post*, including online editions where available. Using this source set, we replicate the original GPR index with a sample correlation of 0.92.[7]

We then use a large language model (LLM), GPT-4.1-nano by OpenAI, to classify each article according to the economic mechanisms it describes. To make the classification rule transparent and reproducible, we formulate the prompt as a structured questionnaire consisting of simple yes-or-no questions, each corresponding to a supply or demand mechanism. An article is assigned to a given shock type if it is classified as describing at least one mechanism mapping to that type. A single article may be assigned to both shock types if its content describes both supply- and demand-side mechanisms; the resulting indices are therefore overlapping subindices of the Caldara-Iacoviello index and need not add up. For each shock type, we construct the corresponding index by dividing the number of classified articles by the total number of articles published in the underlying newspaper corpus and multiplying by 100. This normalization follows the construction of the Caldara–Iacoviello index.

The prompt consists of four parts (see Table 1). The first part defines the task, to classify the article purely by its macroeconomic impact on the United States, and provides concise definitions of canonical negative supply and negative demand shocks. The second part contains the questionnaire. The supply questions ask whether the article explicitly mentions rising costs or prices in the United States, or supply disruptions affecting the U.S. economy, as a consequence of the event. The demand questions analogously ask whether the article explicitly mentions declining U.S. consumer, business, or investor confidence, or weakening demand for goods and services. Explicitly linking each mechanism to the geopolitical event helps distinguish event-driven effects from unrelated economic developments. The third part instructs the model to return its answers in JSON format, facilitating structured data processing. The fourth part appends the text of the article to be classified, under the heading *news article*. The first three parts are identical across articles, meaning the classification rule itself does not vary. We set the sampling temperature to zero, so that the classification is as reproducible as the model permits.

For each yes-or-no answer, the model also reports a coarse self-assessed likelihood that its answer is correct, on a three-point scale: high, medium, or low.[8] In our baseline specification, we retain only classifications rated as high likelihood; including medium-

[7]The Factiva Analytics database does not cover three of the ten sources used in the original GPR index: the *Financial Times*, the *Los Angeles Times*, and the *New York Times*.

[8]We deliberately ask for the term 'likelihood' rather than 'confidence', since the latter already appears in the demand questions, where confidence refers to consumer, business, or investor sentiment. We restrict the scale to three coarse categories because a language model cannot be expected to provide a finely calibrated probability; the field is therefore not interpreted as a calibrated probability but as a coarse label that flags classifications the model itself treats as ambiguous.

Table 1: Questionnaire for identifying supply and demand geopolitical risk indices

You classify news articles about geopolitical events purely by their macroeconomic impact on the United States.

Canonical shocks:
negative_supply = raises production costs or reduces available output
negative_demand = reduces aggregate demand

INSTRUCTIONS:
Read the news article. Base your answer only on the article text. Do not use outside knowledge. Focus only on effects relevant to the United States economy. Think step by step internally. Do NOT reveal your reasoning. Answer each question with exactly "yes" or "no" (lowercase, no punctuation). For each question, also rate the likelihood that your answer is correct as exactly "high", "medium", or "low" (lowercase, no punctuation). Return only the JSON, no extra text, no trailing commas.

QUESTIONS:
negative_supply:
S1. Does the article explicitly mention disruptions of supplies affecting the U.S. economy because of the event?
S2. Does the article explicitly mention rising costs or prices for the U.S. economy because of the event?

negative_demand:
D1. Does the article explicitly mention falling U.S. consumer, business, or investor confidence because of the event?
D2. Does the article explicitly mention falling U.S. consumer, business, or investor demand for goods and services because of the event?

OUTPUT JSON SCHEMA:

```
{
  "supply_disruption": "yes|no",
  "supply_disruption_likelihood": "high|medium|low",
  "cost_increase": "yes|no",
  "cost_increase_likelihood": "high|medium|low",
  "confidence_decline": "yes|no",
  "confidence_decline_likelihood": "high|medium|low",
  "demand_decline": "yes|no",
  "demand_decline_likelihood": "high|medium|low"
}
```

NEWS ARTICLE:

likelihood classifications leaves the classification largely unchanged (see Appendix A).

Under the baseline classification, 3.0% of articles describe only a supply mechanism, 5.2% only a demand mechanism, and 0.3% both, while the remaining 91.5% describe neither mechanism. The small overlap indicates that supply and demand mechanisms are rarely described in the same article, even though the classification permits both. The large residual, in turn, reflects the combination of two design choices. The Caldara–Iacoviello index is topic-defined and therefore selects the broader universe of articles concerning geopolitical risk, consistent with its purpose. Our classification is more restrictive: it retains only articles that explicitly link a geopolitical event to one of the specified mechanisms affecting the U.S. economy and for which the model assigns high likelihood to that classification.

Robustness The baseline prompt is deliberately restrictive: an article is classified only if it explicitly mentions a relevant supply- or demand-side mechanism. This rule-based design limits the scope for an unconstrained interpretation generated by the model. We assess robustness to alternative questionnaire formulations and across language models. The resulting indices correlate with the baseline at 0.95 or above throughout. Additionally, we also aggregate the different questionnaires into ensemble indices, either by assigning an article to a shock type whenever a supermajority of questionnaires does so, or by weighting each article by the share of questionnaires that classify it. Every such ensemble correlates with the baseline index at 0.988 or above. The details are in Appendix A.

3.2 The supply and demand GPR indices over time

Figure 2 plots the two mechanism-specific indices together with the Caldara-Iacoviello GPR index. Both the supply and the demand GPR index in the upper panel fluctuate around 100 in normal times and spike sharply around individual geopolitical episodes. Importantly, however, they spike around *different* episodes, and the distinction tracks the economic character of the underlying events. Conflicts that interrupt the flow of energy and goods are especially visible in the supply index: the Iraqi invasion of Kuwait and the Gulf War, the annexation of Crimea, Russia's full-scale invasion of Ukraine, and the recent Middle East conflict. Events that primarily affect confidence and spending without an immediate commodity-supply component are more prominent in the demand index, most clearly the 9/11 attacks and, later, the Iraq War or the US withdrawal from Afghanistan.

What makes this distinction informative is that we do not assign individual events to shock types ex ante. The language model reads each article and records whether the text explicitly describes a supply mechanism (rising costs, higher prices, or disruptions to the supply of goods) or a demand mechanism (weakening confidence or curtailed spending). The resulting time-series pattern therefore emerges from the textual classification itself. Energy- and trade-disrupting conflicts are especially visible in the supply index, while confidence-related events are more prominent in the demand index, providing an intuitive validation of the classification.

The contrast is particularly clear for the September 2001 attacks and Russia's full-scale invasion of Ukraine in February 2022, two major geopolitical risk episodes that the press framed in markedly different terms. Reporting after 9/11 emphasized weakening spending:

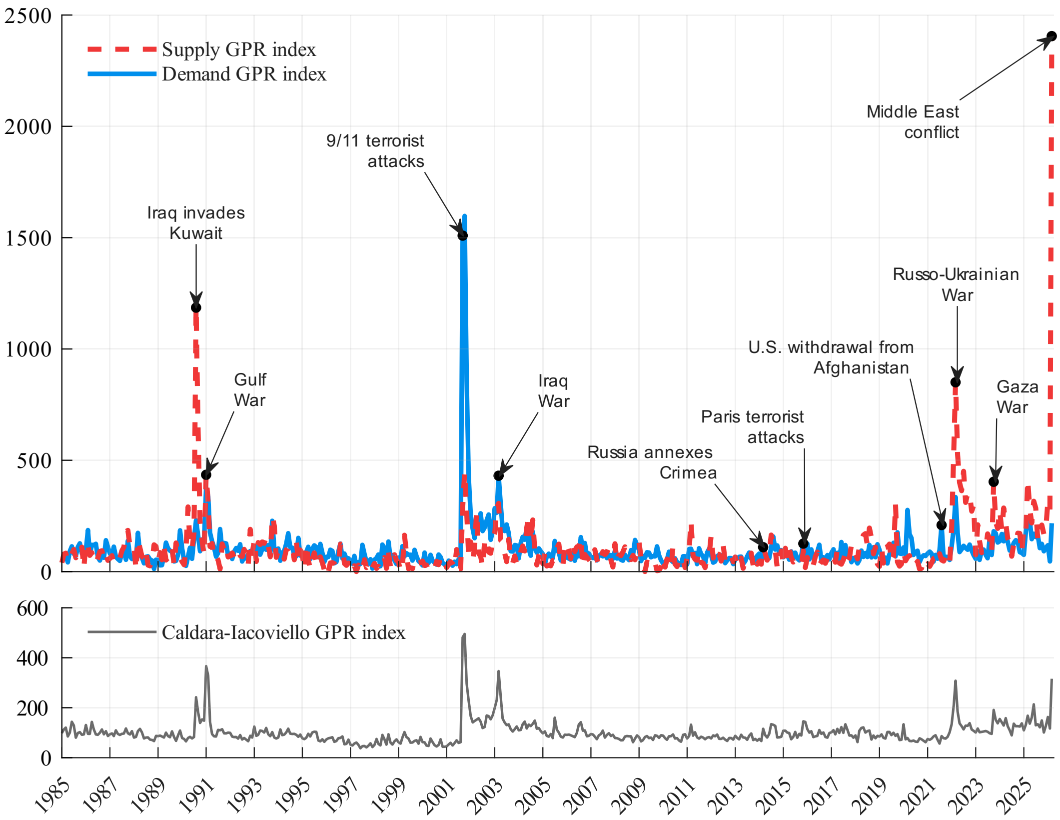


Figure 2: Geopolitical risk indices: Supply, demand, and Caldara–Iacoviello

Notes: The upper panel shows the supply and demand GPR indices obtained from the mechanism-based textual classification, with selected geopolitical events annotated. The lower panel shows the GPR index of Caldara and Iacoviello (2022). All indices are normalized to a mean of 100 over the sample period 1985:1–2026:3.

"Retail stores suffered through one of their weakest Septembers in decades." The Washington Post, 12 October 2001.

Reporting after the invasion of Ukraine instead emphasized commodity prices and supply disruptions:

"Ukraine crisis: commodities prices surge as stock markets slump." The Guardian, 24 February 2022.

"Ukraine War Means Another Supply Shock to Global Economy, the Last Thing It Needs." The Wall Street Journal, 25 February 2022.

The Caldara-Iacoviello GPR index in the lower panel of Figure 2 cannot reveal this distinction. There, 9/11, the Gulf War, the invasion of Ukraine, and the recent Middle East conflict all appear as prominent spikes: 9/11 the largest, the others within roughly a factor of two. The index ranks these episodes only by the intensity of geopolitical-risk coverage; it does not reveal whether the underlying news points primarily to a demand or a supply disturbance. The mechanism-specific indices recover precisely this distinction. Innovations in the Caldara–Iacoviello index therefore combine different structural shocks, and treating them as a single shock leaves the distinction unrecovered.

3.3 Empirical setup

We estimate the dynamic effects of the supply and demand GPR indices in a Bayesian vector autoregression. This subsection describes the data and the empirical model in turn.

Data. Our sample is monthly and spans January 1986 to March 2026. Alongside the GPR index of interest, the VAR includes a standard set of US macroeconomic and financial variables: real GDP, two measures of consumer prices, the short-term interest rate, the ten-year Treasury yield, the S&P 500, and an equity volatility index.

We include two price measures: core PCE, which excludes food and energy, and broad PCE, the overall index. This is because the two shock types transmit to prices through different components and at different speeds. A negative supply shock raises prices most strongly through the volatile food and energy items contained in broad PCE, whereas a negative demand shock works first through the more slowly adjusting core. Including both measures lets the data reveal this difference rather than imposing a single notion of 'the' price response; throughout, we treat core PCE as the primary price measure, as it isolates the price response least mechanically tied to the energy content directly affected by geopolitical events.

Real GDP, available only at quarterly frequency, is interpolated to monthly frequency using the method proposed by Chow and Lin (1971); the short-term interest rate and the volatility index are spliced across, respectively, the zero-lower-bound period and a methodological change in the volatility index. All series are drawn from Haver Analytics and described in more detail in Appendix B.

Empirical model. We estimate a Bayesian VAR with a truncated COVID-19 volatility correction. This approach allows us to exploit the full sample, including the period of the Russo-Ukrainian war since the annexation of Crimea in 2014. The COVID-19 pandemic generated unprecedented volatility in macroeconomic time series, and a growing body of research shows that these extreme observations can distort VAR-based analyses, including both structural inference and forecasting; see, among others, Lenza and Primiceri (2022), Carriero et al. (2024), and Schorfheide and Song (2024).

To address this, we adopt a parsimonious correction that mitigates the influence of pandemic-era outliers while maintaining comparability to previous studies that use standard homoskedastic VAR frameworks in the context of geopolitical risk. Specifically, we follow Lenza and Primiceri (2022) and rescale the covariance matrix for observations within the pandemic period.[9] This variance correction attenuates the disproportionate impact of the pandemic's volatility spikes while preserving the underlying information relevant for inference.

We estimate a VAR of the form

$$y_t = C + \sum_{l=1}^{p} B_l\, y_{t-l} + s_t u_t, \quad u_t \sim N(0, \Sigma),$$

where y_t is an $n \times 1$ vector of endogenous variables at time $t = 1, \ldots, T$, C a vector of constants, B_l coefficient matrices of size $n \times n$, and u_t an $n \times 1$ vector of reduced-form residuals. The auxiliary variable s_t scales the residual covariance matrix and equals one prior to the pandemic. From the onset of the pandemic, denoted t^*, the model assumes $s_{t^*} = \bar{s}_0$, $s_{t^*+1} = \bar{s}_1$, $s_{t^*+2} = \bar{s}_2$, and

$$s_{t^*+j} = 1 + (\bar{s}_2 - 1)\rho^{j-2},$$

where $[\bar{s}_0, \bar{s}_1, \bar{s}_2, \rho]$ is a vector of unknown coefficients of the covariance adjustment.

A key concern in our setting is that an unrestricted volatility correction may downweight observations after the Russian invasion of Ukraine. In particular, the decay parameter ρ determines how quickly the volatility scaling reverts to unity. Under a standard specification, the posterior of ρ implies that the correction remains active well into 2022 and 2023, a period that is central to our analysis. Accordingly, we impose a truncation on the volatility adjustment by setting $s_t = 1$ for all t beyond 18 months after the onset of the pandemic (i.e., from October 2021 onward). This ensures that observations from the invasion period onward are no longer downweighted in the likelihood, preserving the identifying variation in the GPR indices that is relevant to our analysis.[10]

For estimation, we use prior distributions for the VAR coefficients of the conjugate Normal-Inverse Wishart type, defined through a set of hyperparameters. In particular, we employ the standard Minnesota prior (Litterman, 1986), but relax the assumption of a random walk for the GPR index by setting the first lag to 0.8. As a prior for $\bar{s}_0$, $\bar{s}_1$, and $\bar{s}_2$, we assume a Pareto distribution with scale and shape equal to one, which allows for potentially large increases in the reduced-form residuals over the pandemic. For ρ, the model imposes a Beta prior, which we calibrate to the standard values proposed in the literature. The lag length p is set to twelve. Impulse responses are computed over a 12-month horizon; we report median estimates together with 68% posterior credible bands.

[9] An alternative approach would be to simply omit the extreme observations associated with the onset of the pandemic, specifically March, April, and May 2020. However, for the analysis of GPR shocks, excluding these months is particularly problematic, as they precede the Russian invasion of Ukraine and can contain valuable pre-crisis information.

[10] Under this specification, the posterior median of ρ is 0.83, indicating that the correction still captures the gradual normalization of volatility during the pandemic period.

We estimate three specifications, one for each GPR index: the supply, the demand, and the original Caldara–Iacoviello index. In each case, the index is ordered first in the VAR. As discussed in Section 2, this internal-instrument ordering identifies the structural impulse responses to the proxy under the assumption that its innovations are orthogonal to other contemporaneous shocks. For the Caldara–Iacoviello index, we additionally implement sign-restriction identification within the model, as a benchmark against which to assess the ex ante classification approach.

3.4 Macroeconomic effects of supply and demand GPR shocks

We now turn to the macroeconomic effects of the two shocks. We first present the impulse responses and then assess their structural interpretation using the coherence criterion of Section 2.2.

3.4.1 Impulse responses

Figures 3 and 4 show the responses to a one-standard-deviation innovation in the supply and demand GPR index. We plot median responses together with 68% posterior credible bands over a twelve-month horizon. The indices themselves respond persistently, decaying over roughly six months. The two shocks trace out distinct and theory-consistent patterns, differing in the sign of the price response, the monetary-policy reaction, and the timing of the real effects.

The supply GPR shock displays the defining feature of an adverse supply disturbance: output and prices move in opposite directions. GDP declines on impact and continues to fall, reaching close to -0.1% by the end of the first year. Prices rise over the same horizon: core PCE increases gradually to about 0.05% and broad PCE to about 0.06%, the latter rising both faster on impact and by more, consistent with the commodity-sensitive food and energy components that it includes and the core index strips out. The policy rate is essentially unchanged, consistent with a central bank that looks through the supply-driven rise in prices rather than tightening into the contemporaneous fall in output. The ten-year Treasury yield rises quickly, by up to 0.06 percentage points, and remains elevated. Equity prices fall persistently, by roughly 1.0% at the one-year horizon, while the VIX jumps by about 2.4% on impact and returns to baseline within a few months. The response builds rather than peaking on impact: both the decline in output and the rise in prices deepen over the year, consistent with the gradual pass-through of higher input costs and the propagation of supply disruptions.

The demand GPR shock differs on precisely the dimension that separates the two shock types. Output and prices now move together: GDP falls on impact, its strongest response, around -0.06%, and the decline fades over the following months, while core PCE also falls on impact, by close to 0.015%, before drifting back up. Broad PCE declines

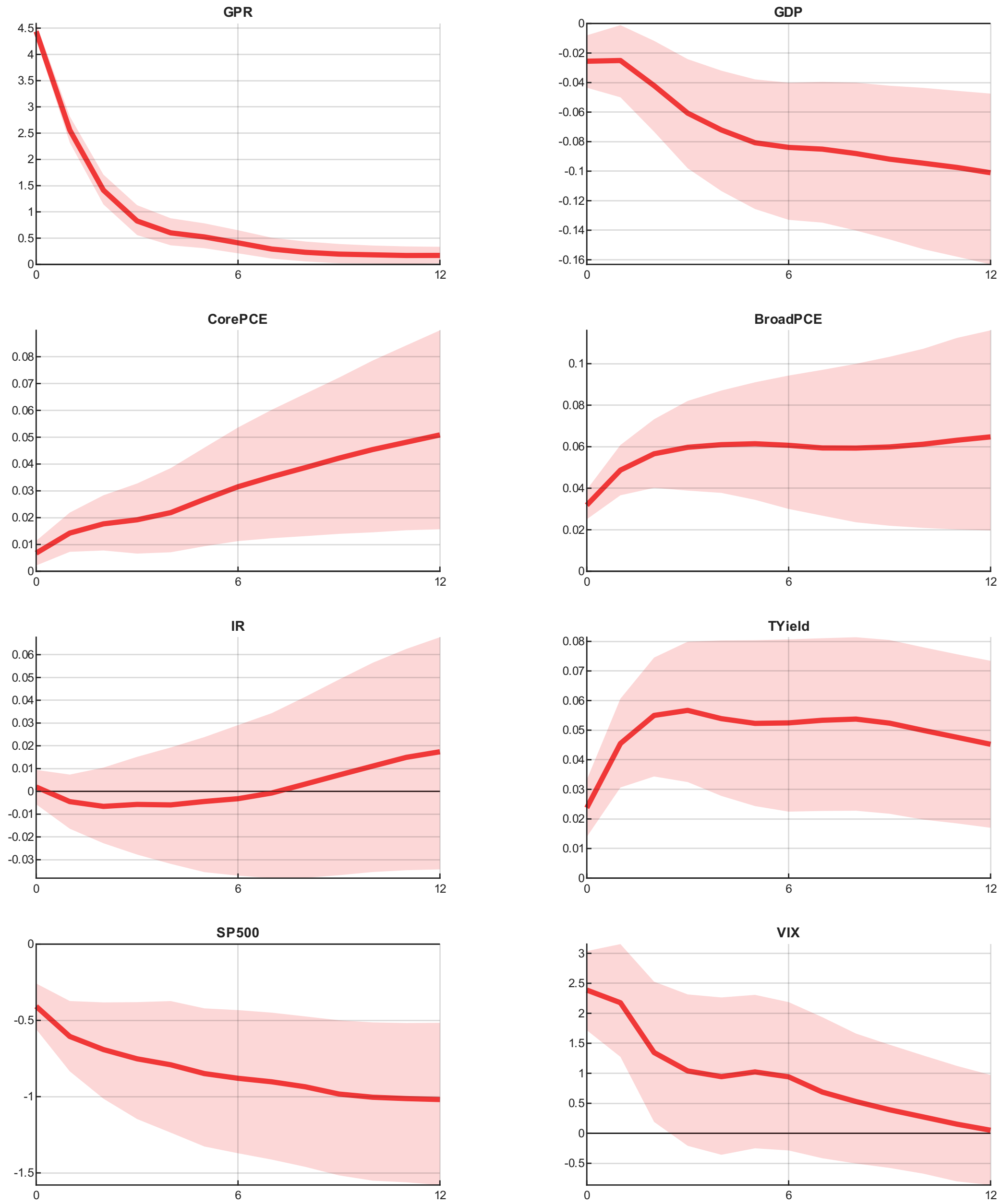


Figure 3: Supply GPR index

Notes: Impulse responses to a one-standard-deviation innovation in the supply GPR index. The solid line is the posterior median and the shaded area the 68% highest posterior density band. The horizon is in months. GPR, GDP, core PCE, broad PCE, the S&P 500, and the VIX are in percent; the policy rate (IR) and the 10-year Treasury yield (TYield) are in percentage points. Variable definitions, data sources, and the identifying assumptions are detailed in Section 3.3.

with a delay, reaching a trough of roughly -0.023% only around the fourth month. This contemporaneous comovement, output and prices both falling, is the signature of an adverse demand shock; we evaluate coherence on impact, where the theoretical prediction is arguably the sharpest.[11] The monetary and financial responses sharpen the distinction.

[11]The criterion itself is not tied to contemporaneous responses: the signature in Section 2.2 is a set of restrictions that may involve any horizon. We restrict attention to contemporaneous responses because theory delivers its arguably sharpest predictions there, whereas responses at longer horizons may

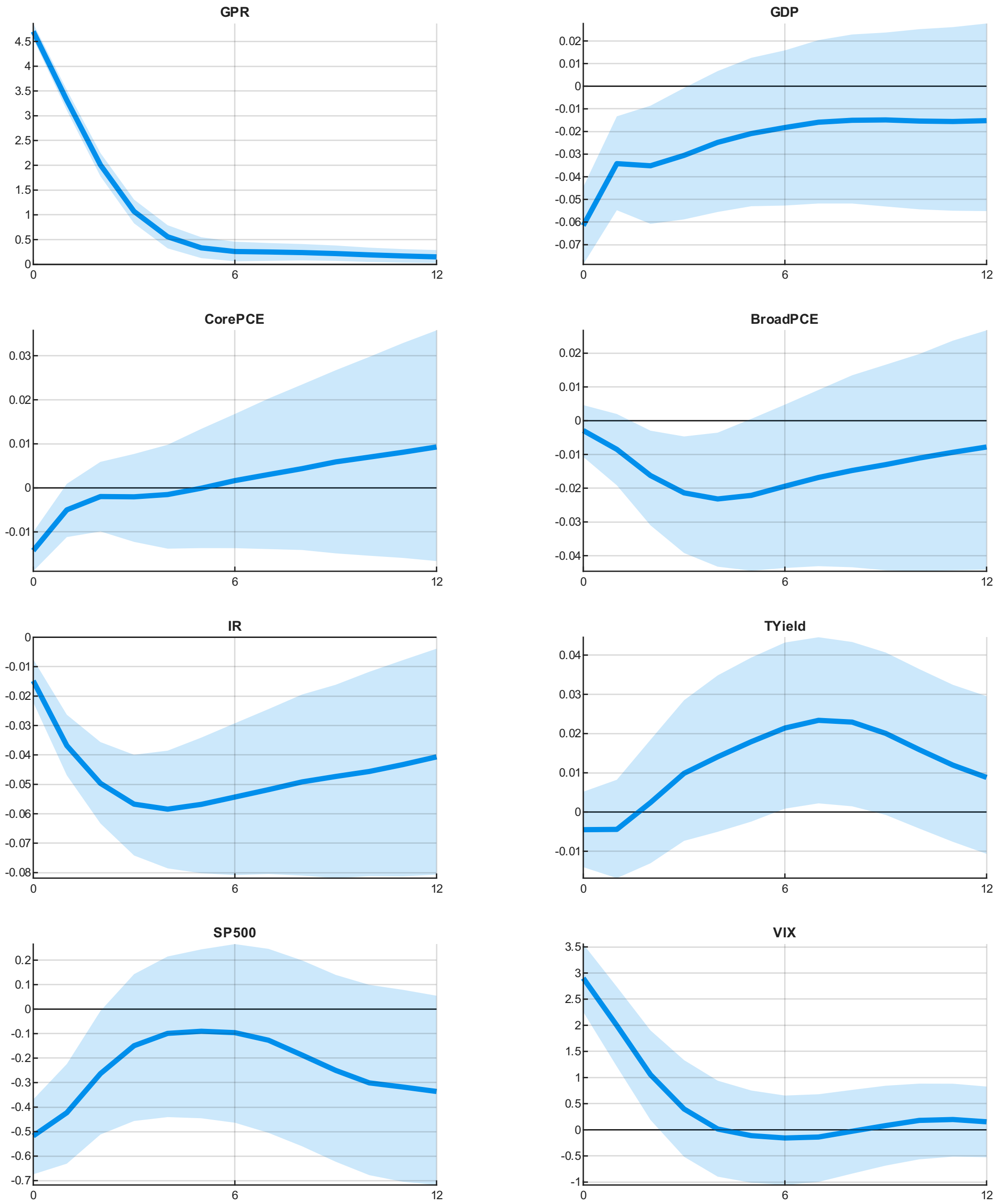


Figure 4: Demand GPR index

Notes: Impulse responses to a one-standard-deviation innovation in the demand GPR index. The solid line is the posterior median and the shaded area the 68% highest posterior density band. The horizon is in months. GPR, GDP, core PCE, broad PCE, the S&P 500, and the VIX are in percent; the policy rate (IR) and the 10-year Treasury yield (TYield) are in percentage points. Variable definitions, data sources, and the identifying assumptions are detailed in Section 3.3.

The policy rate now eases on impact, falling by up to about 0.06 percentage points, in contrast to the look-through response under the supply shock. This is the expected difference: a demand shock moves output and prices in the same direction and calls for unambiguous easing, whereas a supply shock confronts the central bank with a tradeoff. The ten-year yield shows no response on impact and moves only later in the horizon.

increasingly reflect the endogenous policy reaction.

Equity prices fall by about 0.5%, smaller than under the supply shock and short-lived rather than persistent, and the VIX jumps more on impact, by about 2.9%, with markedly tighter credible bands, so the immediate rise in financial uncertainty is estimated more sharply. In contrast to the supply shock, whose effects build over the year, the demand shock is front-loaded: strongest on impact and fading thereafter.

The two shocks therefore differ on exactly the dimensions that separate demand from supply: the sign of the price response, the monetary policy reaction, and the timing of the real effects. The following coherence analysis allows us to formally quantify the extent to which these patterns accord with the assigned signatures.

3.4.2 Coherence validation

Our classification assigned each index a shock type without reference to the macroeconomic data: an article was labeled supply or demand by the mechanism it described, not by how output or prices subsequently moved. The coherence criterion of Section 2.2 asks whether the impulse responses are consistent with the assigned shock type. The criterion is informative precisely because the responses are free to disagree with the assigned type. In contrast, a within-model identification that imposes the signature satisfies it by construction, so the diagnostic cannot provide evidence. An external classification leaves the responses unrestricted, so agreement, where it occurs, is evidence.

We find that the impulse responses to supply and demand GPR shocks are consistent with their assigned signatures. Conditioning the impulse responses on the assigned signature leaves them essentially unchanged: the credible bands computed from only the draws consistent with the signature nearly coincide with those from the full posterior (Figure 5). In the plane of the contemporaneous output and price responses, the posterior draws concentrate in the quadrant implied by the assigned shock type: output down with prices up for the supply index, output and prices down together for the demand index (Figure 6). The diagnostic (BF) ranges from 18.1 to 330.7 for the supply shock and from 5.5 to 3,832.4 for the demand shock (Table 2). All values exceed the threshold for positive evidence, so both shocks are coherent with their assigned types across all tested signatures. The lower demand BF arises on broad prices: the median impact response is negative, but the credible set includes zero on impact and excludes it only from the third month onward, too late for the contemporaneous signature.[12]

Each shock is also evaluated against the signature of the opposing type. This placebo

[12] For the supply GPR index, the multivariate diagnostic (BF = 330.7) exceeds the diagnostics on GDP with core PCE (BF = 18.1) and with broad PCE (BF = 40.2): the multivariate signature is less likely to be satisfied by chance, so it is the more demanding test. For the demand GPR index the ordering reverses when comparing the multivariate diagnostic with the GDP–core PCE diagnostic (BF = 57.5 against 3,832.4). This is because two of the additionally restricted responses (broad PCE and the Treasury yield) are correctly signed but quantitatively close to zero. The diagnostic nonetheless remains strong.

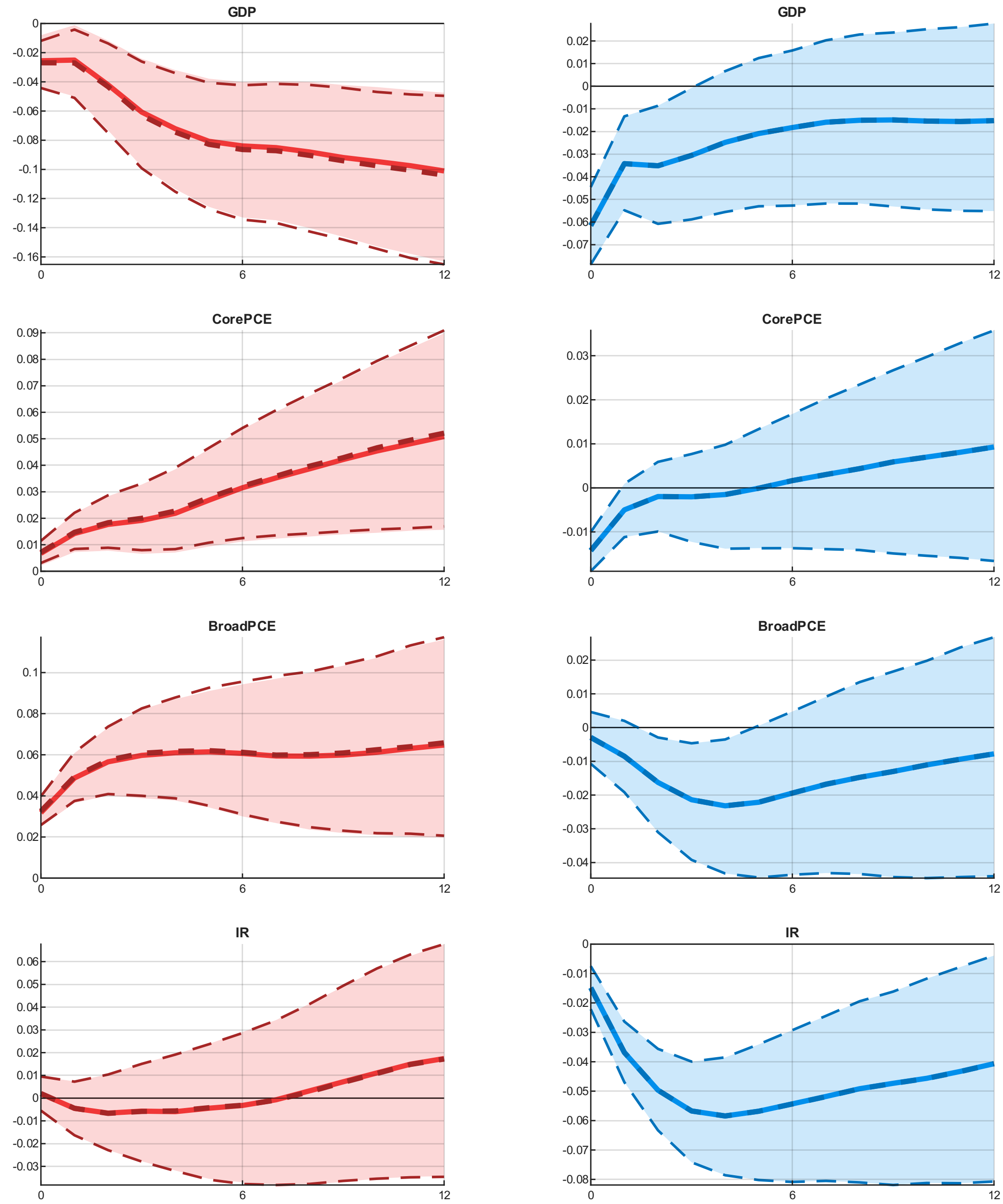
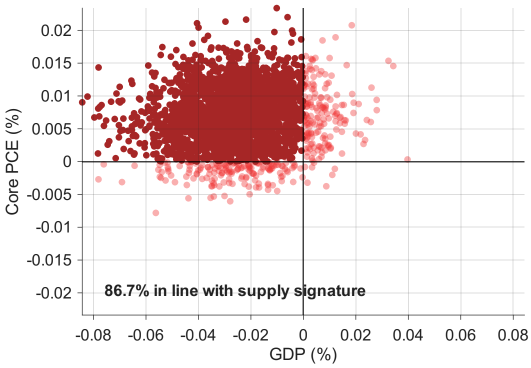


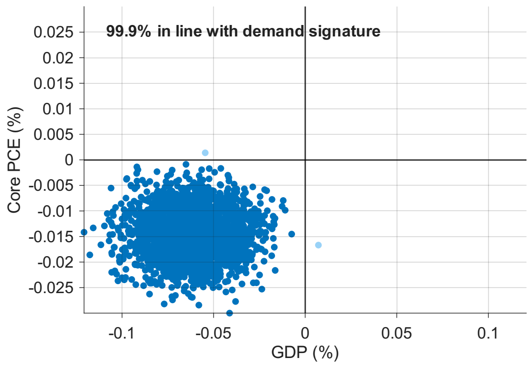
Figure 5: Coherence of the supply and demand GPR shocks evaluated on contemporaneous GDP and core PCE

Notes: Impulse responses of GDP, core PCE, broad PCE, and the policy rate to a one-standard-deviation innovation in the supply GPR index (left, red) and the demand GPR index (right, blue), identified by ordering the index first in the VAR. Solid lines and shaded areas are the posterior median and 68% highest posterior density band, computed over the full posterior. The dashed lines are the corresponding median (thick) and 68% band (thin) computed over only the coherent draws, i.e. the posterior draws whose contemporaneous responses satisfy the assigned signature: GDP falling with core PCE rising for the supply shock, GDP and core PCE falling together for the demand shock. GDP, core PCE, and broad PCE are in percent; the policy rate (IR) is in percentage points. The horizon is in months. Variable definitions and the identifying assumptions are detailed in Section 3.3; the coherence criterion is defined in Section 2.2.

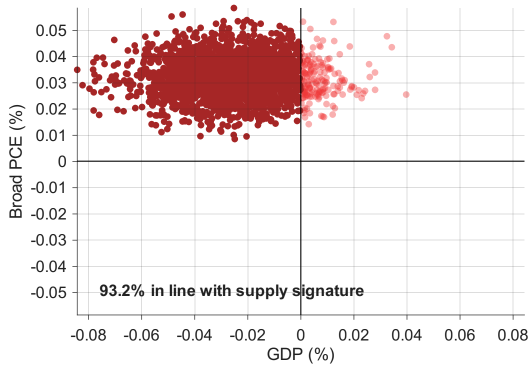
test asks whether the classification truly distinguishes the two shock types: if it does, the shock recovered from each proxy should not only be coherent with its own signature



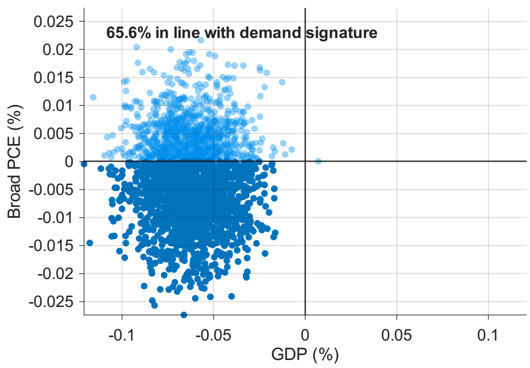
(a) Supply GPR: Core PCE



(b) Demand GPR: Core PCE



(c) Supply GPR: Broad PCE



(d) Demand GPR: Broad PCE

Figure 6: Contemporaneous GDP and price responses for supply and demand GPR shocks

Notes: Each point is one posterior draw of the contemporaneous responses of GDP (horizontal axis) and the price level (vertical axis) to a one-standard-deviation innovation in the supply GPR index (left column) or the demand GPR index (right column). The price measure is core PCE in the top row and broad PCE in the bottom row. Dark colored dots mark draws that satisfy a signature: dark red the supply signature (GDP<0, prices>0) and dark blue the demand signature (GDP<0, prices<0). The percentage in each panel is the share of draws consistent with the shock's assigned signature: the coherence share C reported in Table 2. The signature is evaluated against the price measure shown, so the same shock is read against core PCE in the top row and broad PCE in the bottom row. Variable definitions and the identifying assumptions are detailed in Section 3.3.

but also fail to be coherent with the other one, i.e. the diagnostic should provide no evidence for the wrong interpretation. This is what we find, with no evidence for the opposing signature. For the shock recovered from the supply GPR index, coherence with the demand signature falls to 6.4% for core PCE, well below the prior benchmark of 24.6% (BF $=0.2$), and the broad-PCE and multivariate diagnostics yield a coherence of 0.0 (BF $=0.0$). The demand shock mirrors this pattern ($C = 0.2\%$ and 0.0% on core PCE and in the multivariate version, BF $=0.0$; BF $=1.5$ on broad PCE, again reflecting the near-zero impact response of broad PCE). Each shock is thus coherent with its assigned type, while the data provide no support for the opposing interpretation; the assignment is not interchangeable.

Table 2: Coherence validation for supply and demand GPR indices

		C (%)	C_0 (%)	BF
Panel A: Supply GPR index				
Supply signature	GDP, core PCE	86.7	26.5	18.1
	GDP, broad PCE	93.2	25.3	40.2
	Multivariate	85.8	1.8	330.7
Demand signature	GDP, core PCE	6.4	24.6	0.2
	GDP, broad PCE	0.0	25.8	0.0
	Multivariate	0.0	1.5	0.0
Panel B: Demand GPR index				
Supply signature	GDP, core PCE	0.0	26.5	0.0
	GDP, broad PCE	34.4	25.3	1.5
	Multivariate	0.0	1.8	0.0
Demand signature	GDP, core PCE	99.9	24.6	3832.4
	GDP, broad PCE	65.6	25.8	5.5
	Multivariate	46.7	1.5	57.5

Notes: C is the share of posterior draws whose contemporaneous impulse responses match the stated signature. C_0 is the corresponding share under the prior benchmark, and BF $= [C/(1-C)]/[C_0/(1-C_0)]$ measures how far likelihood updating moves the odds toward the signature relative to that benchmark (the coherence criterion); by Kass and Raftery (1995), BF above 3 indicates positive, above 20 strong, and above 150 very strong evidence. A shock's signature is the sign its contemporaneous responses should take under the assigned type: under a supply shock GDP falls while prices rise; under a demand shock GDP and prices fall together. The first two rows of each block restrict GDP together with one price measure, core or broad PCE; the *multivariate* signature restricts GDP, both price measures, the 10-year yield (supply > 0, demand < 0), the S&P 500 (< 0), and the VIX (> 0), leaving the policy rate unrestricted. Each index is evaluated against both its assigned signature (matched) and the opposing one (placebo).

Robustness Finally, as a robustness check, we repeat the analysis for ensemble indices that assign an article to a shock type whenever at least five of the eight questionnaire formulations do so (Appendix A). The impulse responses lie close to the baseline medians throughout, and the coherence validation reproduces the baseline pattern: the shocks recovered from the ensemble indices are coherent with their assigned types on the same scale as the baseline, and the placebo evaluation again finds no evidence for the opposing interpretation.

3.5 Comparison with the Caldara-Iacoviello GPR index

The mechanism-specific indices draw the supply–demand distinction at the measurement stage, before any restriction is placed on the model. The GPR index of Caldara and Iacoviello (2022) carries no such information; it measures the intensity of geopolitical risk without regard to any mechanism. It is therefore natural to ask whether the separation our classification provides could instead be extracted from the Caldara-Iacoviello GPR

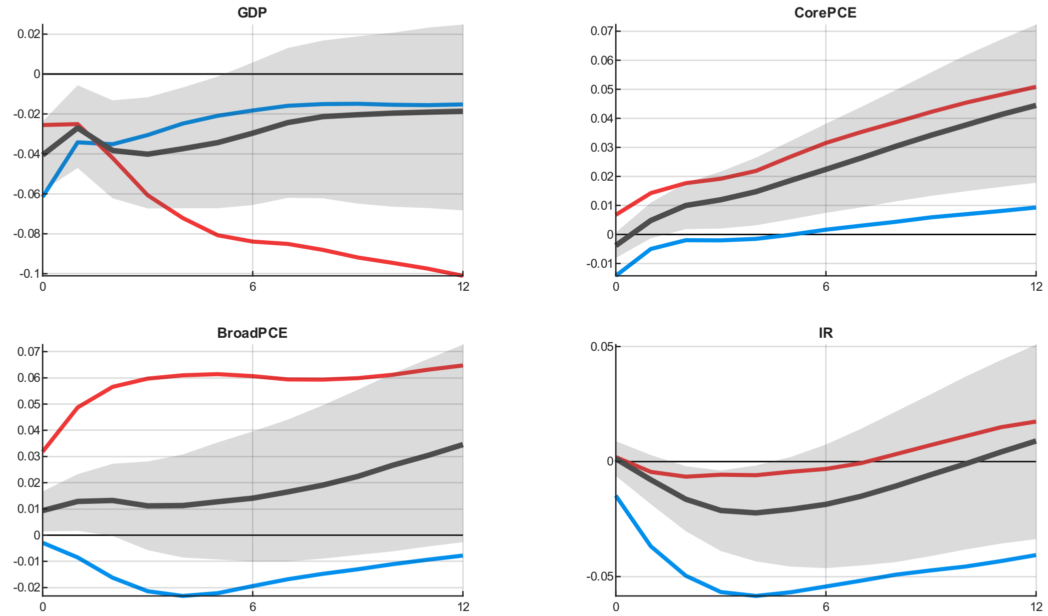


Figure 7: Recursively identified Caldara-Iacoviello GPR shock compared with the supply and demand GPR shocks

Notes: Impulse responses of GDP, core PCE, broad PCE, and the policy rate to a one-standard-deviation innovation in the Caldara–Iacoviello GPR index, identified by ordering the index first in the VAR. The black line is the posterior median and the grey shaded area the 68% highest posterior density band. Overlaid for comparison are the posterior medians of the responses to the supply GPR index shock (red) and the demand GPR index shock (blue), each estimated in its own VAR and reproduced from Figures 3 and 4. GDP, core PCE, and broad PCE are in percent; the policy rate (IR) is in percentage points. The horizon is in months. Variable definitions and the identifying assumptions are detailed in Section 3.3.

index alone, through identification carried out within the model. We follow the two standard routes, a recursive ordering and within-model sign restrictions, and find that neither separates the shocks: applied to this GPR index, both return a single disturbance that blends the supply and demand responses rather than isolating either.

Recursive identification. We first identify a single geopolitical risk shock recursively, ordering the Caldara–Iacoviello GPR index first. Figure 7 reports the resulting impulse responses (black, with 68% bands) and overlays the posterior medians of the supply GPR shock (red) and the demand GPR shock (blue), each estimated in its own VAR and normalized to a one-standard-deviation innovation in its own index. In every panel, the response to the Caldara–Iacoviello shock lies between the two mechanism-specific responses, and the blend is not uniform across variables. The output decline is demand-like: pronounced on impact and easing over the horizon, rather than deepening gradually as under the supply shock. The price responses are supply-like in sign but muted in size: core and broad PCE both rise, yet remain below the supply shock's path throughout, pulled toward zero by the demand component.[13] The policy rate eases mildly and with a delay, between the pronounced easing under the demand shock and the near-flat response under the supply shock. The recursively identified shock thus replicates neither mechanism-specific response; it averages them.[14]

Both signatures require output to fall, so the output response cannot separate the two interpretations. The type must be read off prices. Which signature the diagnostic favors, however, depends on the price measure. Figure 8 plots the same posterior draws of the contemporaneous responses against core PCE and against broad PCE (panel (a) and (b), respectively); the output responses are common to both panels, but the price responses are not. On core PCE the draws lie mostly below zero and the diagnostic favors the demand signature ($C = 79.6\%$, BF $= 12.0$); on broad PCE the same draws lie mostly above zero and it favors the supply signature ($C = 87.1\%$, BF $= 20.0$), with no evidence for the opposing signature in each case (Table 3, Panel A). The recursively identified GPR shock thus has no stable structural identity: its apparent type depends on the price measure used. The mechanism-specific shocks, by contrast, keep their assigned identity under both measures (Table 2). The shock recovered from the index is therefore a mixture of the supply and demand disturbances rather than either in isolation. The distinction between them is not encoded in the index, and recursive identification cannot recover what the data do not contain.

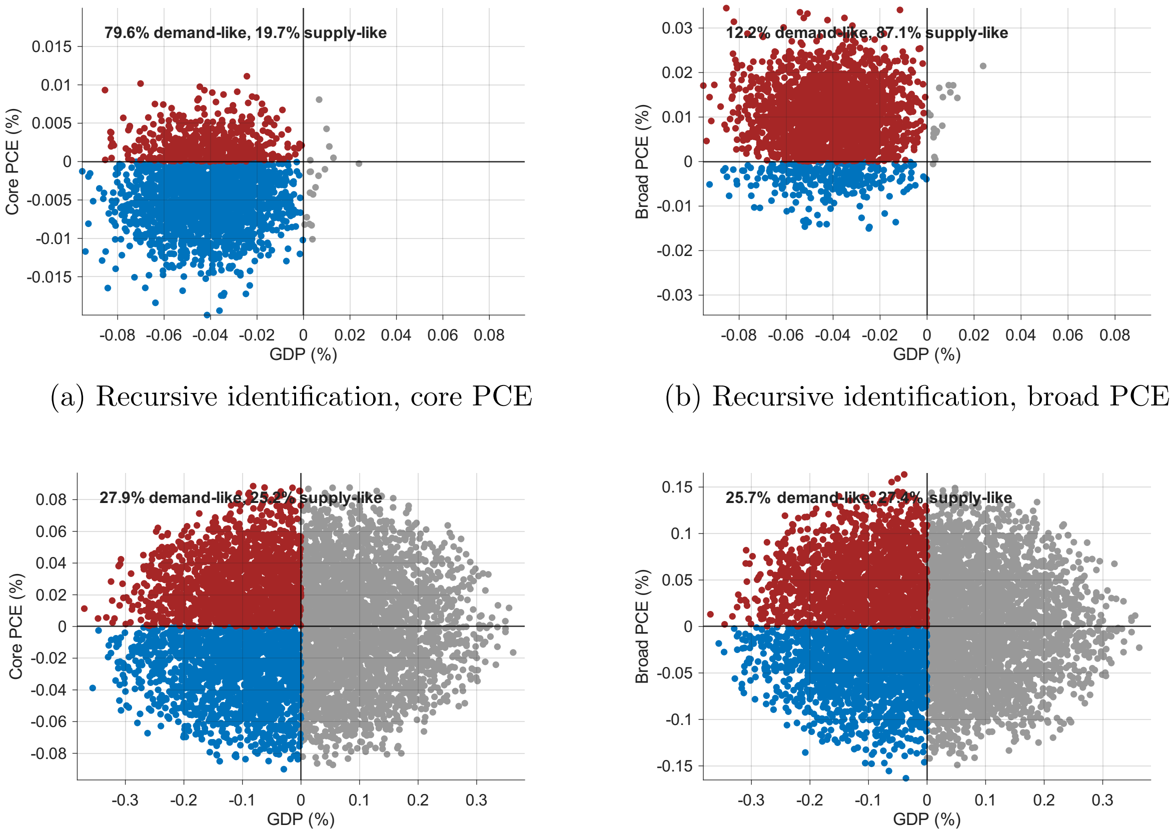
Sign restrictions. A natural objection is that the separation could instead be imposed within the model, through sign restrictions. Following Rubio-Ramírez et al. (2010) and Arias et al. (2018), we draw orthogonal rotation matrices Q from the Haar measure and retain those whose implied contemporaneous response of the GPR index is positive; this base cloud imposes no restriction on the macroeconomic responses. We then identify two shocks from the index by retaining the draws whose contemporaneous responses match the supply signature (output falling, core PCE rising) and, separately, the demand signature (output and core PCE both falling) alongside the requirement that the index rises. Each shock is thus identified on its own, as in the mechanism-specific exercise (Figure 9).

The two identified shocks using sign restrictions coincide on every response except the one the restrictions force apart: output falls in both, as it is restricted to, and core PCE is forced apart, while the unrestricted responses (broad PCE, the policy rate, the long yield, equity prices, and the VIX) are largely indistinguishable, their credible bands overlapping throughout.[15] The only systematic difference between the supply and demand shocks is

[13]Estimates of the effects of a recursively identified GPR shock are reported across a large set of countries by Caldara et al. (2026) and across the euro area by Bondarenko et al. (2026).

[14]Appendix Figure 14 reports the complementary signature-conditional decomposition: recomputing the responses over only the draws that satisfy the supply or the demand signature.

[15]Appendix Figure 15 overlays the two sign-restricted shocks on the base cloud of all rotations consistent with a rising GPR index. Outside the restricted responses, the two sets of bands largely coincide with each other and with the unrestricted cloud, indicating that the restrictions do not materially sharpen



(a) Recursive identification, core PCE

(b) Recursive identification, broad PCE

(c) Sign restrictions, core PCE

(d) Sign restrictions, broad PCE

Figure 8: Contemporaneous GDP and price responses to a Caldara–Iacoviello GPR shock: recursive and sign-restriction identification.

Notes: Each point is a posterior draw of the contemporaneous responses of GDP (horizontal axis) and a price measure (vertical axis) to a shock identified from the Caldara–Iacoviello GPR index; the left column uses core PCE, the right column broad PCE. The top row identifies the shock by a recursive ordering with the index first; the bottom row identifies it by sign restrictions on the index, with gray points showing the full set of rotations consistent with GPR> 0. Because the sign-restriction step admits many admissible rotations per posterior draw, the bottom-row clouds are randomly thinned to 10% of the rotated draws for legibility; the reported shares are computed from the full set of draws. Red points are draws consistent with the supply signature (GDP< 0, price> 0), blue points with the demand signature (GDP< 0, price< 0). Under the recursive ordering the draws shift across the price-response axis with the price measure used, so the same shock reads as demand-like against core PCE and supply-like against broad PCE. Under sign restrictions the draws form a single contiguous cloud that the restriction merely bisects; the supply and demand regions are adjacent halves of one cloud rather than separated clusters, and the share of draws in each region stays close to its prior benchmark (Table 3).

the one the restrictions impose.

Figure 8 (panels (c) and (d)) shows why. The rotations consistent with a rising GPR index form a single cloud spanning all four quadrants of the contemporaneous output–price plane. The supply and demand regions (the upper-left and lower-left quadrants) are adjacent parts of this cloud rather than separated clusters, and the restriction merely

inference about the remaining responses. This lack of separation is related to the shock-masquerading problem in Wolf (2020): linear combinations of other structural shocks can satisfy sign restrictions intended to isolate a particular disturbance.

Table 3: Coherence validation of the Caldara–Iacoviello GPR index

		C (%)	C_0 (%)	BF
Panel A: Cholesky identification				
Supply signature	GDP, core PCE	19.7	26.5	0.7
	GDP, broad PCE	87.1	25.3	20.0
Demand signature	GDP, core PCE	79.6	24.6	12.0
	GDP, broad PCE	12.2	25.8	0.4
Panel B: Sign-restriction identification				
Supply signature	GDP, core PCE	25.2	25.5	1.0
	GDP, broad PCE	27.4	24.6	1.2
Demand signature	GDP, core PCE	27.9	24.1	1.2
	GDP, broad PCE	25.7	25.0	1.0

Notes: The table reports the coherence criterion for shocks identified recursively and through sign restrictions from the Caldara–Iacoviello GPR index. C is the share of posterior draws whose contemporaneous responses match the stated signature; C_0 is the corresponding share under the prior benchmark; and BF $= [C/(1-C)]/[C_0/(1-C_0)]$ measures how far likelihood updating moves the odds toward the signature relative to that benchmark (the coherence criterion). By Kass and Raftery (1995), BF above 3 indicates positive, above 20 strong, and above 150 very strong evidence. A shock's signature is the sign its contemporaneous responses should take under the assigned type: under a supply shock GDP falls while prices rise; under a demand shock GDP and prices fall together. In Panel A the shock is identified by a recursive (Cholesky) ordering with the index ordered first; in Panel B, coherence is computed over the cloud of rotations whose contemporaneous GPR response is positive, with no restriction on the macroeconomic responses; C is the share of that cloud whose contemporaneous GDP and price responses match the stated signature.

bisects it: the share of draws in each region stays close to its prior benchmark, on core PCE (panel c) as on broad PCE (panel d).

The coherence criterion provides no evidence that sign-restriction identification can achieve a separation of the supply and demand shock.[16] The coherence with each signature stays within a few percentage points of its prior benchmark on both price measures, leaving every BF diagnostic between 1.0 and 1.2 (Table 3, Panel B).[17] This is what the criterion records when identification imposes the signature rather than leaving the data free to reject it or not: a shock restricted to satisfy a signature cannot provide evidence for it. Sign restrictions thus do not recover the supply–demand distinction from the Caldara–Iacoviello index; they impose it, and outside the restricted responses the data neither distinguish the two shocks nor corroborate the pattern imposed. The separation drawn by the ex ante classification at the measurement stage cannot be reconstructed within the model from a covariance structure that does not encode it.

[16]The coherence diagnostic is computed over the base cloud and evaluated against each signature on either price measure, exactly as for the recursively identified shock.

[17]The limited updating within the sign-restricted set is consistent with Baumeister and Hamilton (2015), who show that when structural parameters remain set identified, the asymptotic posterior is proportional to the prior within the identified set.

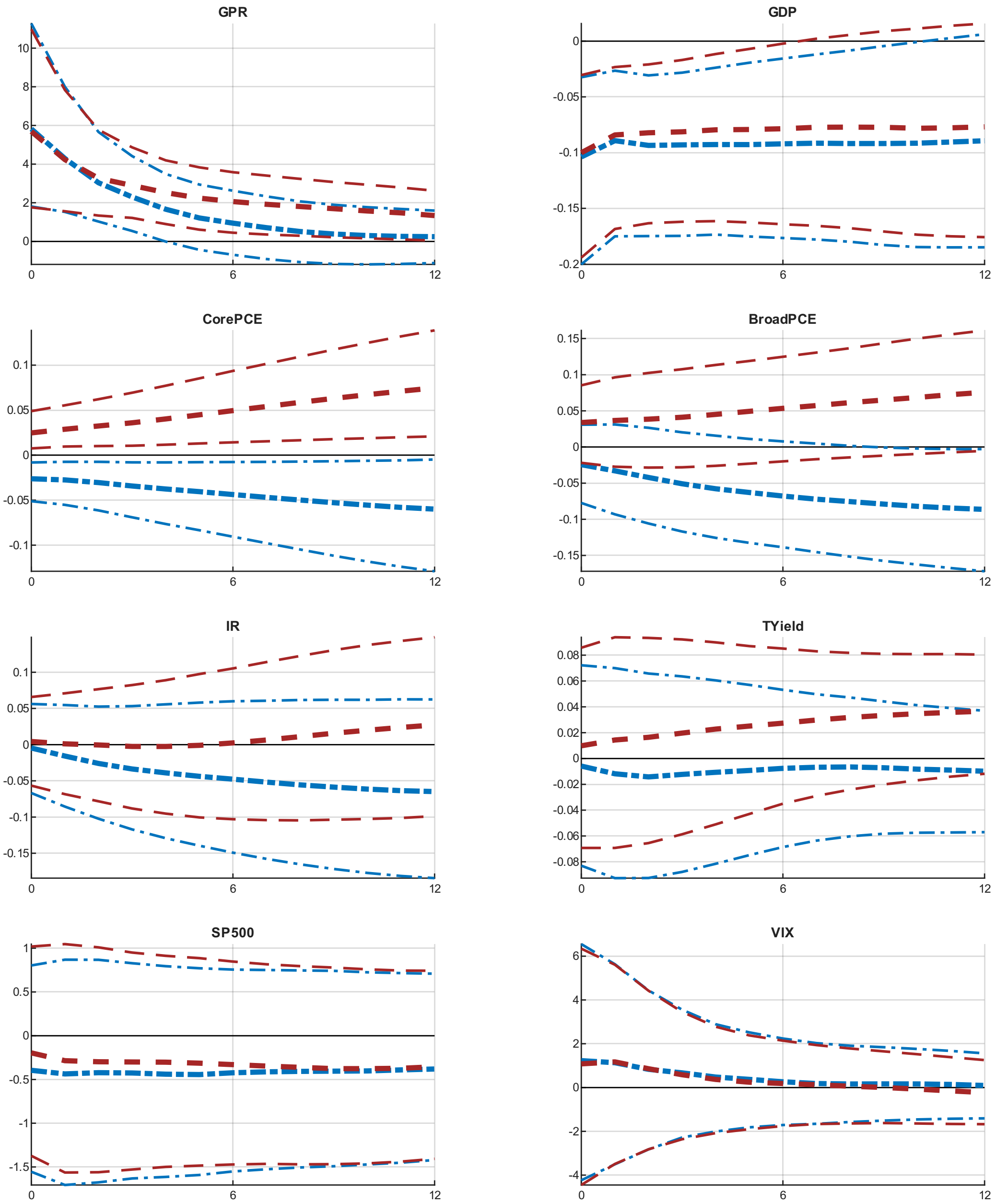


Figure 9: Sign-restriction identification on the Caldara–Iacoviello GPR index: supply and demand shocks

Notes: Impulse responses to two shocks identified from the Caldara–Iacoviello GPR index by sign restrictions on the contemporaneous responses. The supply shock (red) imposes a rise in the aggregate GPR index, a fall in GDP, and a rise in core PCE; the demand shock (blue) imposes a rise in the GPR index and a fall in both GDP and core PCE. All remaining responses (broad PCE, the policy rate, the 10-year yield, the S&P 500, and the VIX) are left unrestricted. Thick dash(-dotted) lines are the posterior medians and thin dash(-dotted) lines the 68% highest posterior density bands. GPR, GDP, core PCE, broad PCE, the S&P 500, and the VIX are in percent; the policy rate (IR) and the 10-year yield (TYield) are in percentage points. The horizon is in months. Variable definitions and the sign-restriction scheme are detailed in Section 3.3.

4 Conclusion

In empirical macroeconomics, it has become commonplace to construct shock proxies independent of any particular model: narrative records, high-frequency surprises, and text-based indices serve as proxies for monetary policy shocks, tax changes, economic policy uncertainty, geopolitical risk, and climate-related risk, among others. Validation has not kept pace: once a proxy is constructed, its structural interpretation typically rests on the quality of the external information and the assumptions used to isolate it, rather than on evidence from the estimated model. This paper combines identification with validation: shocks are identified outside the model, by classifying events according to the economic mechanism they describe, and their structural interpretation is validated inside it. Because the theoretical signature of a shock is used neither to construct the proxy nor to recover the shock, it remains available as a testable implication; our coherence criterion measures the posterior share of impulse-response draws consistent with the assigned signature, benchmarked against the corresponding prior share.

Applied to geopolitical risk, the framework yields mechanism-specific supply and demand GPR indices from the article corpus underlying the Caldara–Iacoviello index. In a Bayesian VAR, the two indices generate distinct, theory-consistent responses, differing in the price response, the monetary policy reaction, and the timing of real effects. The shocks recovered from them are strongly coherent with their assigned types, and the placebo evaluation finds no evidence for the opposing interpretation. Innovations to the Caldara–Iacoviello index itself behave as a mixture: under a recursive ordering, the structural identity of the shock switches with the price measure used to read it, and sign restrictions satisfy their signature by construction, so the criterion registers no evidence beyond the prior. We find that the distinction the classification draws at the measurement stage cannot be recovered within the model after the fact.

The mechanism-specific indices are of independent interest. They indicate whether a geopolitical event is transmitting through costs and supply or through confidence and spending; channels with opposite implications for inflation. Because the classification uses no macroeconomic data, the indices can be constructed in real time: when a geopolitical event breaks, the split between supply- and demand-side coverage is observable within days, well before output and price data reveal the transmission. The indices may thus help central banks gauge the inflationary consequences of geopolitical events as they unfold.

The coherence criterion itself applies to any externally identified shock for which theory implies a response signature, and several applications suggest themselves. Narrative and high-frequency monetary policy surprises are natural candidates: their structural interpretation is central to a large literature and can now be evaluated rather than assumed. Topic-defined text indices of uncertainty and risk are another, since the question of which structural shocks their innovations blend arises there in the same form as for geopolitical risk. Statistically identified shocks, finally, are recovered unlabeled; the criterion supplies the outside information that their economic interpretation requires.

References

Acosta, M. (2026). The perceived causes of monetary policy surprises. Working Paper.

Antolín-Díaz, J. and J. Rubio-Ramírez (2018). Narrative sign restrictions for SVARS. *American Economic Review 108*, 2802–2829.

Arias, J. E., J. F. Rubio-Ramírez, and D. F. Waggoner (2018). Inference based on structural vector autoregressions identified with sign and zero restrictions: Theory and applications. *Econometrica 86*, 685–720.

Aruoba, S. B. and T. Drechsel (2026). Identifying monetary policy shocks: A natural language approach. *American Economic Journal: Macroeconomics*. Forthcoming.

Baker, S. R., N. Bloom, and S. J. Davis (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics 131*, 1593–1636.

Baumeister, C. and J. D. Hamilton (2015). Sign restrictions, structural vector autoregressions, and useful prior information. *Econometrica 83*, 1963–1999.

Blanchard, O. and D. Quah (1989). The dynamic effects of aggregate demand and supply disturbances. *American Economic Review 79*, 655–673.

Bondarenko, Y., N. Kang, V. Lewis, M. Rottner, and Y. Schüler (2026). Geopolitical risk in the euro area: Measurement and transmission. *Deutsche Bundesbank Discussion Paper 14/2026*.

Bondarenko, Y., V. Lewis, M. Rottner, and Y. Schüler (2024). Geopolitical risk perceptions. *Journal of International Economics 152*, 104005.

Caldara, D., S. Conlisk, M. Iacoviello, and M. Penn (2026). Do geopolitical risks raise or lower inflation? *Journal of International Economics 159*, 104188.

Caldara, D. and M. Iacoviello (2022). Measuring geopolitical risk. *American Economic Review 112*, 1194–1225.

Carriero, A., T. E. Clark, M. Marcellino, and E. Mertens (2024). Addressing COVID-19 outliers in BVARs with stochastic volatility. *The Review of Economics and Statistics 106*, 1403–1417.

Chow, G. and A.-L. Lin (1971). Best linear unbiased interpolation, distribution, and extrapolation of time series by related series. *The Review of Economics and Statistics 53*, 372–375.

Clayton, C., A. Coppola, M. Maggiori, and J. Schreger (2025). Geoeconomic pressure. *NBER Working Paper 34020*.

Doh, T., D. Song, and S.-K. Yang (2026). Deciphering federal reserve communication via text analysis of alternative FOMC statements. *American Economic Journal: Macroeconomics 18*, 379–412.

Engle, R. F., S. Giglio, B. Kelly, H. Lee, and J. Stroebel (2020). Hedging climate change news. *The Review of Financial Studies 33*, 1184–1216.

Gertler, M. and P. Karadi (2015). Monetary policy surprises, credit costs, and economic activity. *American Economic Journal: Macroeconomics 7*, 44–76.

Gürkaynak, R., B. Sack, and E. Swanson (2005). Do actions speak louder than words? the response of asset prices to monetary policy actions and statements. *International Journal of Central Banking 1*, 55–93.

Handlan, A. (2022). Text shocks and monetary surprises: Text analysis of FOMC statements with machine learning. Working paper.

Hassan, T. A., S. Hollander, L. van Lent, and A. Tahoun (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics 134*, 2135–2202.

Iacoviello, M. and J. Tong (2026). The AI-GPR index: Measuring geopolitical risk using artificial intelligence. Mimeo.

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*, 773–795.

Kuttner, K. (2001). Monetary policy surprises and interest rates: Evidence from the Fed funds futures market. *Journal of Monetary Economics 47*, 523–544.

Kwon, B., T. Park, P. Rungcharoenkitkul, and F. Smets (2025, October). Parsing the pulse: Decomposing macroeconomic sentiment with LLMs. BIS Working Papers 1294, Bank for International Settlements.

Lenza, M. and G. E. Primiceri (2022). How to estimate a vector autoregression after March 2020. *Journal of Applied Econometrics 37*, 688–699.

Litterman, R. (1986). Forecasting with Bayesian vector autoregressions – Five years of experience. *Journal of Business and Economic Statistics 4*, 25–38.

Lumbanraja, A., S. Mouabbi, E. Passari, and A. Rousset Planat (2026). Beyond oil: The origins of commodity price fluctuations. CEPR Discussion Paper 21244, Centre for Economic Policy Research.

Mertens, K. and M. Ravn (2013). The dynamic effects of personal and corporate income tax changes in the United States. *American Economic Review 103*, 1212–47.

Mohr, C. and C. Trebesch (2025). Geoeconomics. *Annual Review of Economics 17*.

Plagborg-Møller, M. and C. K. Wolf (2021). Local projections and VARs estimate the same impulse responses. *Econometrica 89*, 955–980.

Quilis, E. (2018). Temporal disaggregation of economic time series: The view from the trenches. *Statistica Neerlandica 72*, 447–470.

Ramey, V. A. (2011). Identifying government spending shocks: It's all in the timing. *The Quarterly Journal of Economics 126*, 1–50.

Rigobon, R. (2003). Identification through heteroskedasticity. *The Review of Economics and Statistics 85*, 777–792.

Romer, C. D. and D. H. Romer (1989). Does monetary policy matter? a new test in the spirit of Friedman and Schwartz. *NBER Macroeconomics Annual 4*, 121–170.

Romer, C. D. and D. H. Romer (2004). A new measure of monetary shocks: Derivation and implications. *American Economic Review 94*, 1055–1084.

Rubio-Ramírez, J., D. Waggoner, and T. Zha (2010). Structural vector autoregressions: Theory of identification and algorithms for inference. *The Review of Economic Studies 77*, 665–696.

Schorfheide, F. and D. Song (2024). Real-time forecasting with a (standard) mixed-frequency VAR during a pandemic. *International Journal of Central Banking 20*, 275–320.

Sims, C. (1980). Macroeconomics and reality. *Econometrica 48*, 1–48.

Stock, J. and M. Watson (2012). Disentangling the channels of the 2007-09 recession. *Brookings Papers on Economic Activity 43*, 81–156.

Stock, J. H. and M. W. Watson (2018). Identification and estimation of dynamic causal effects in macroeconomics using external instruments. *The Economic Journal 128*, 917–948.

Uhlig, H. (2005). What are the effects of monetary policy on output? results from an agnostic identification procedure. *Journal of Monetary Economics 52*, 381–419.

Wolf, C. K. (2020). SVAR (mis)identification and the real effects of monetary policy shocks. *American Economic Journal: Macroeconomics 12*, 1–32.

Wu, J. and F. Xia (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit and Banking 48*, 253–291.

A Classification robustness

The mechanism classification rests on two implementation choices: the wording of the questionnaire and the language model that answers it. We assess sensitivity along both dimensions.

Questionnaire variation. We vary the baseline questionnaire along three binary dimensions and run all eight resulting combinations (a complete 2^3 factorial design) over the full article corpus. The first factor removes the causal anchor ("because of the event") from each indicator question; without it, the model may attribute background macroeconomic conditions to the article rather than to the geopolitical event it describes. The second factor adds a forward-looking expectations channel, so that an article mentioning effects *expected to* materialize is classified alongside one describing contemporaneous effects. The third factor reverses the order in which the shock definitions and the instructions appear in the prompt. Table 4 summarizes the eight resulting variants. All variants share the baseline output schema, so classifications are directly comparable without post-processing. The full text of all eight questionnaires is available in the replication package.

Table 4: Questionnaire variations

Variant	Prompt configuration			Deviations from baseline
	Causal anchor	Expectations	Definitions first	
Baseline	yes	no	yes	0
Prompt 1	no	no	yes	1
Prompt 2	yes	yes	yes	1
Prompt 3	yes	no	no	1
Prompt 4	no	yes	yes	2
Prompt 5	no	no	no	2
Prompt 6	yes	yes	no	2
Prompt 7	no	yes	no	3

Notes. Each row states the configuration of the three prompt factors. *Causal anchor*: the indicator questions include the phrase "because of the event". *Expectations*: the questions also cover effects expected to materialize, not only contemporaneous ones. *Definitions first*: the canonical shock definitions precede the instructions in the prompt. The baseline configuration is (yes, no, yes); variants are ordered by the number of factors switched relative to it.

The upper panel of Table 5 reports, for each variant, the share of articles that receive the same label as under the baseline questionnaire and the correlation of the resulting monthly index with the baseline index. Under the baseline high-likelihood threshold, agreement ranges from 97.3% to 98.8% for supply classifications and from 93.1% to 96.9% for demand classifications, and the corresponding index correlations range from 0.982 to 0.994 and from 0.957 to 0.988. Demand classifications are consistently more sensitive to the questionnaire than supply classifications, consistent with the more diffuse language through which demand mechanisms (confidence, spending) are described, relative to the concrete vocabulary of costs and supply disruptions. Agreement is uniformly lower once medium-likelihood classifications are included (94.8–98.2% for supply and 91.6–96.5% for demand), which supports restricting the baseline to high-likelihood answers.

Table 5: Sensitivity of the classification to the questionnaire and the model

	Article-level agreement (%)		Index correlation	
Variant	Supply	Demand	Supply	Demand
Questionnaire variation				
Prompt 1	97.8	95.9	0.987	0.978
Prompt 2	98.1	94.7	0.989	0.971
Prompt 3	98.8	96.9	0.994	0.986
Prompt 4	97.3	96.4	0.982	0.983
Prompt 5	98.3	96.8	0.992	0.988
Prompt 6	98.3	93.1	0.993	0.957
Prompt 7	97.3	96.9	0.985	0.987
Model variation				
GPT-4.1-mini	96.5	94.5	0.958	0.954

Notes. Agreement is the share of articles to which the variant assigns the same yes/no label as the baseline; correlation is the Pearson correlation between the resulting monthly indices, computed on the common sample 1985M1–2026M3. Both are reported separately for supply and demand classifications and use the high-likelihood threshold. Questionnaire variants are defined in Table 4 and are classified with the baseline model (GPT-4.1-nano); the model variation applies the baseline questionnaire with GPT-4.1-mini.

Model variation. We repeat the baseline classification with GPT-4.1-mini, holding the questionnaire, the likelihood threshold, and the sampling temperature fixed. The lower panel of Table 5 reports article-level agreement of 96.5% for supply and 94.5% for demand classifications, comparable to what a reformulation of the questionnaire produces. The resulting monthly indices correlate with the baseline at 0.958 and 0.954, falling at the end of the range spanned by the questionnaire variants. The two models thus disagree on a similar share of articles as two questionnaires do, but their disagreements fall more heavily on the quiet months in which no salient geopolitical event dominates coverage, so they displace the index by somewhat more than a reformulation of the questions does.

The above two exercises vary *how* the mechanisms are elicited, not *which* mechanisms.

The supply and demand indices correlate with each other at 0.25 to 0.39, across all 64 pairings of questionnaires and under the alternative model as well. Therefore, the separation between the two indices is not an artifact of the implementation that produces it.

Ensembles. Finally, we aggregate the eight questionnaires into ensemble indices, either by assigning an article to a shock type whenever a supermajority of questionnaires (five or six out of eight) does so, or by weighting each article by the share of questionnaires that classify it. Every such ensemble correlates with the baseline index at 0.988 or above. The interchangeability extends to the recovered shocks. For the five-of-eight ensemble, the impulse responses lie close to the baseline medians throughout (Figures 10–11), and Table 6 repeats the coherence validation: the shocks recovered from the ensemble indices are coherent with their assigned types on the same scale as the baseline, and the placebo evaluation again finds no evidence for the opposing interpretation. We nevertheless build the paper on the single baseline questionnaire: since it is interchangeable with any ensemble, we prefer the object that can be concisely stated in full and easily applied by others. It makes it the more transparent object both for replication and for exposition.

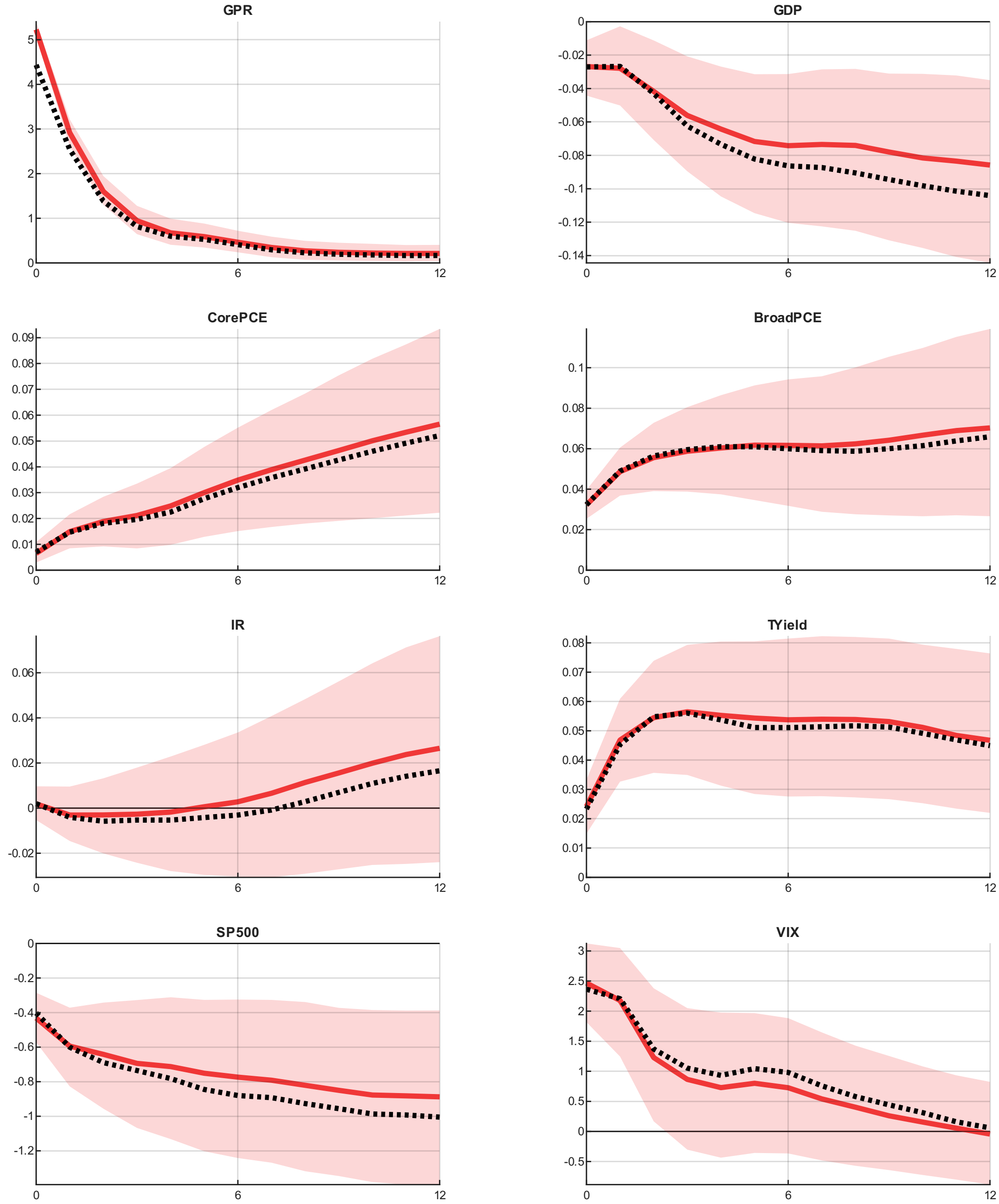


Figure 10: Robustness: Five-of-eight ensemble Supply GPR index

Notes: Impulse responses to a one-standard-deviation innovation in the supply GPR index constructed from the five-of-eight ensemble classification. The solid line is the posterior median and the shaded area the 68% highest posterior density band. The dotted black line is the posterior median of the responses to the supply GPR index constructed from the baseline questionnaire, reproduced from Figure 3. The horizon is in months. GPR, GDP, core PCE, broad PCE, the S&P 500, and the VIX are in percent; the policy rate (IR) and the 10-year Treasury yield (TYield) are in percentage points. Variable definitions, data sources, and the identifying assumptions are detailed in Section 3.3.

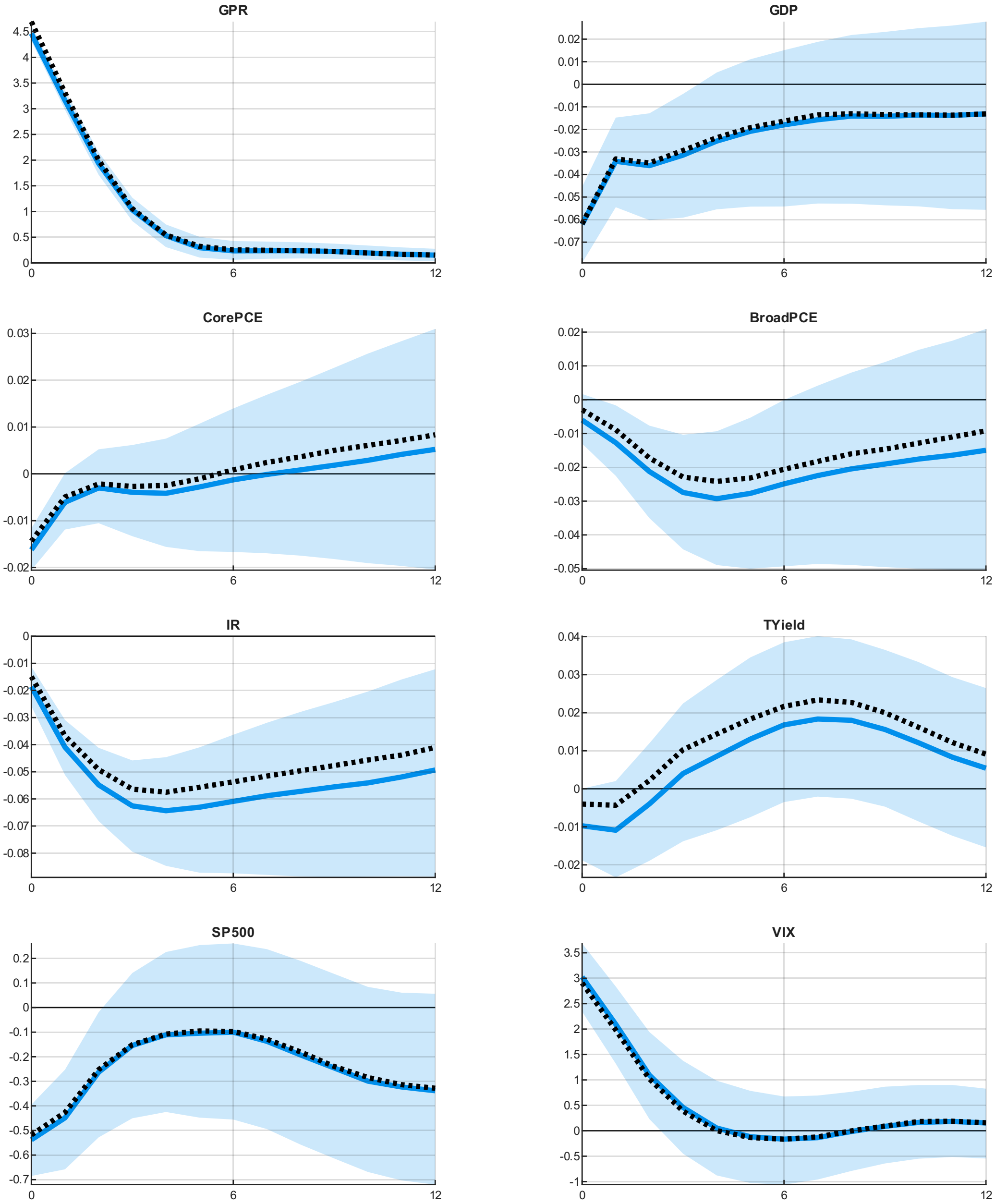


Figure 11: Robustness: Five-of-eight ensemble demand GPR index

Notes: Impulse responses to a one-standard-deviation innovation in the demand GPR index constructed from the five-of-eight ensemble classification. The solid line is the posterior median and the shaded area the 68% highest posterior density band. The dotted black line is the posterior median of the responses to the demand GPR index constructed from the baseline questionnaire, reproduced from Figure 4. The horizon is in months. GPR, GDP, core PCE, broad PCE, the S&P 500, and the VIX are in percent; the policy rate (IR) and the 10-year Treasury yield (TYield) are in percentage points. Variable definitions, data sources, and the identifying assumptions are detailed in Section 3.3.

Table 6: Coherence validation for the five-of-eight ensemble GPR indices

		C (%)	C_0 (%)	BF
Panel A: Supply GPR index				
Supply signature	GDP, core PCE	83.7	26.5	14.2
	GDP, broad PCE	92.0	25.3	34.0
	Multivariate	83.2	1.8	270.2
Demand signature	GDP, core PCE	8.3	24.6	0.3
	GDP, broad PCE	0.0	25.8	0.0
	Multivariate	0.0	1.5	0.0
Panel B: Demand GPR index				
Supply signature	GDP, core PCE	0.0	26.5	0.0
	GDP, broad PCE	22.0	25.3	0.8
	Multivariate	0.0	1.8	0.0
Demand signature	GDP, core PCE	100.0	24.6	30,680.4
	GDP, broad PCE	78.0	25.8	10.2
	Multivariate	66.3	1.5	129.3

Notes: The table repeats the coherence validation for the ensemble indices that assign an article to a shock type whenever at least five of the eight questionnaires do so, retaining only high-likelihood classifications; the estimation and identification are otherwise identical to the baseline. C is the share of posterior draws whose contemporaneous impulse responses match the stated signature. C_0 is the corresponding share under the prior benchmark, and $\text{BF} = [C/(1-C)]/[C_0/(1-C_0)]$ measures how far likelihood updating moves the odds toward the signature relative to that benchmark (the coherence criterion); by Kass and Raftery (1995), BF above 3 indicates positive, above 20 strong, and above 150 very strong evidence. A shock's signature is the sign its contemporaneous responses should take under the assigned type: under a supply shock GDP falls while prices rise; under a demand shock GDP and prices fall together. The first two rows of each block restrict GDP together with one price measure, core or broad PCE; the *multivariate* signature restricts GDP, both price measures, the 10-year yield (supply > 0, demand < 0), the S&P 500 (< 0), and the VIX (> 0), leaving the policy rate unrestricted. Each index is evaluated against both its assigned signature (matched) and the opposing one (placebo).

B Data construction

This appendix documents the construction of all macroeconomic variables used in the empirical analysis. Our sample is monthly and spans January 1986 to March 2026. All variables are sourced from Haver Analytics. Table 7 lists the variables, their abbreviations used throughout the paper, and their Haver codes.

Three variables require additional construction and are documented in the subsections below: real GDP (Chow and Lin (1971) disaggregation to monthly frequency), the short-term interest rate (splice across the zero lower bound), and equity volatility (splice across the VIX/VXO methodology change). GDP, CorePCE, BroadPCE, Equity, and VIX enter the VAR in log levels (multiplied by 100). IR and TYield enter in levels (percent per annum).

Table 7: Variables, abbreviations, and Haver codes

Abbreviation	Description	Haver code
GDP	Real Gross Domestic Product (see B.1)	GDPH@USECON
CorePCE	PCE less Food & Energy Chain Price Index	JCXFEBM@USECON
BroadPCE	PCE Chain Price Index	JCBM@USECON
IR	Effective Federal Funds Rate (see B.2)	FFED@USECON
	Wu-Xia Shadow Federal Funds Rate	FFEDSHDW@USECON
TYield	10-Year Treasury Constant Maturity Yield	FCM10@USECON
Equity	S&P 500 Composite	SP500@USECON
VIX	CBOE Volatility Index (see B.3)	SPVIX@USECON
	CBOE S&P 100 Volatility Index (VXO)	SPVXO@USECON
Chow-Lin disaggregation inputs (see B.1):		
	Industrial Production Index	IP@USECON
	Real Personal Consumption Expenditures	CBHM@USECON
	Real Personal Income excl. Transfer Receipts	YPXTPHM@USECON
	All Employees: Total Nonfarm	LANAGRA@USECON
	Real Manufacturing & Trade Sales	TSTH@USECON

B.1 Monthly real GDP via Chow-Lin disaggregation

Real Gross Domestic Product is published at quarterly frequency. We disaggregate it to monthly frequency using the Chow and Lin (1971) temporal disaggregation method, which estimates monthly values that are consistent with the quarterly aggregate and use information from correlated monthly indicators. We use five monthly indicators: industrial production, real personal consumption expenditures, real personal income excluding transfer receipts, total nonfarm employment, and real manufacturing and trade sales. The disaggregation is implemented using the MATLAB toolbox provided by Quilis (2018).

B.2 Interest rate during the zero lower bound

Our interest rate combines the Effective Federal Funds Rate outside the zero lower bound with the Wu and Xia (2016) Shadow Federal Funds Rate during ZLB periods. The splice points correspond to official FOMC decisions on the target range. The first ZLB episode runs from December 2008 to December 2015, the second from March 2020 to March 2022.

B.3 Equity volatility index across the VIX/VXO transition

Our equity volatility measure is the CBOE Volatility Index. The official VIX series, constructed under CBOE's post-2003 methodology, is available from January 1990 onward, with the pre-2003 portion retrospectively constructed by CBOE using the current methodology. For the period from January 1986 to December 1989, when VIX is not available, we extend the series using the CBOE S&P 100 Volatility Index, CBOE's original volatility index calculation. To ensure continuity at the splice point, we scale the VXO series by the ratio of VIX to VXO observed in the overlapping sample. Our sample begins in January 1986 to coincide with the start of VXO availability.

C Additional results

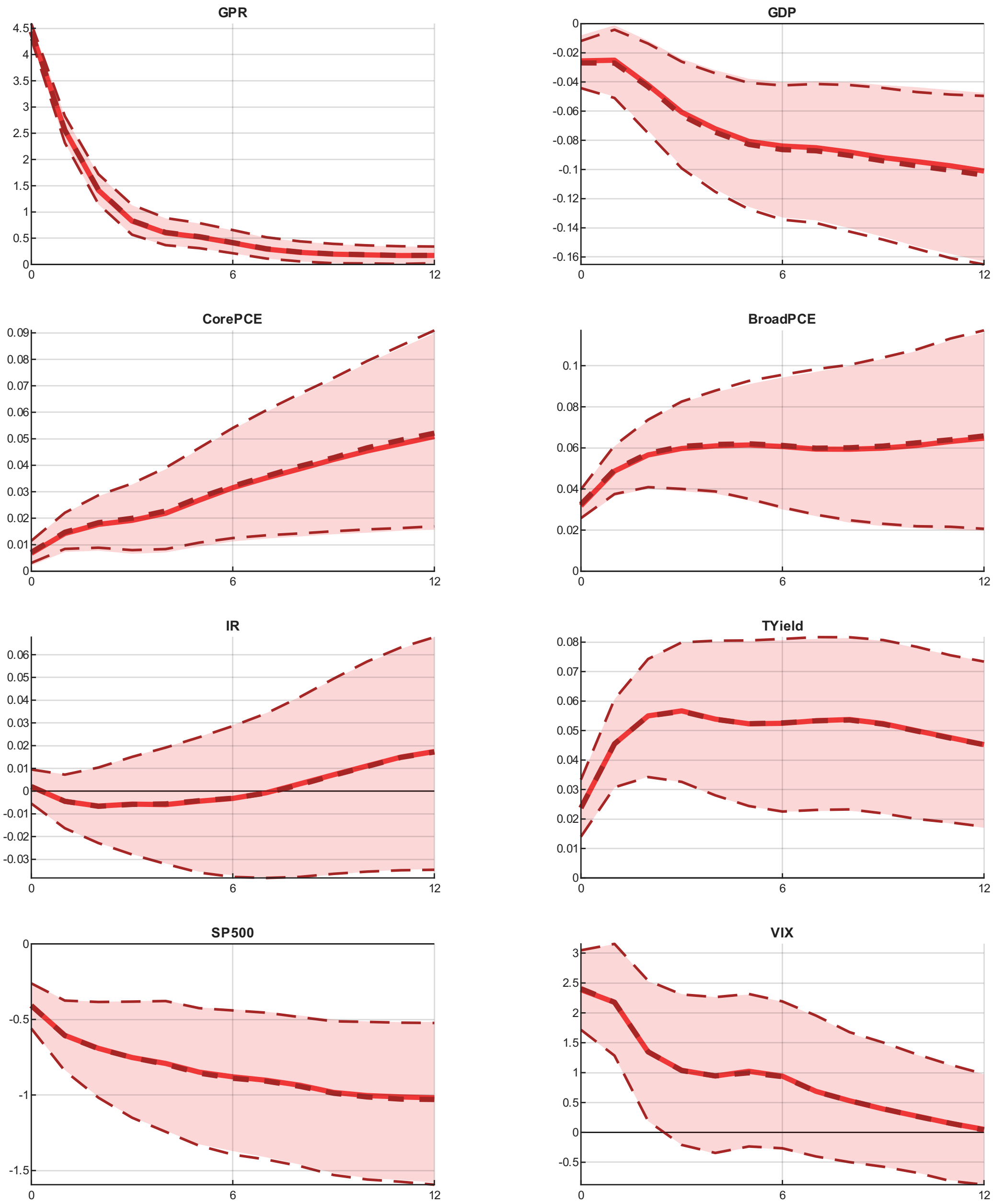


Figure 12: Coherence of the supply GPR shock evaluated on contemporaneous GDP and core PCE

Notes: Impulse responses to a one-standard-deviation innovation in the supply GPR index, identified by ordering the index first in the VAR. Solid lines and shaded areas are the posterior median and 68% highest posterior density band, computed over the full posterior. The dashed lines are the corresponding median (thick) and 68% band (thin) computed over only the coherent draws, i.e. the posterior draws whose contemporaneous responses satisfy the assigned signature: GDP falling with core PCE rising. GPR, GDP, core PCE, broad PCE, S&P 500, and VIX are in percent; the policy rate (IR) and 10-year yield (TYield) are in percentage points. The horizon is in months. Variable definitions and the identifying assumptions are detailed in Section 3.3; the coherence criterion is defined in Section 2.2.

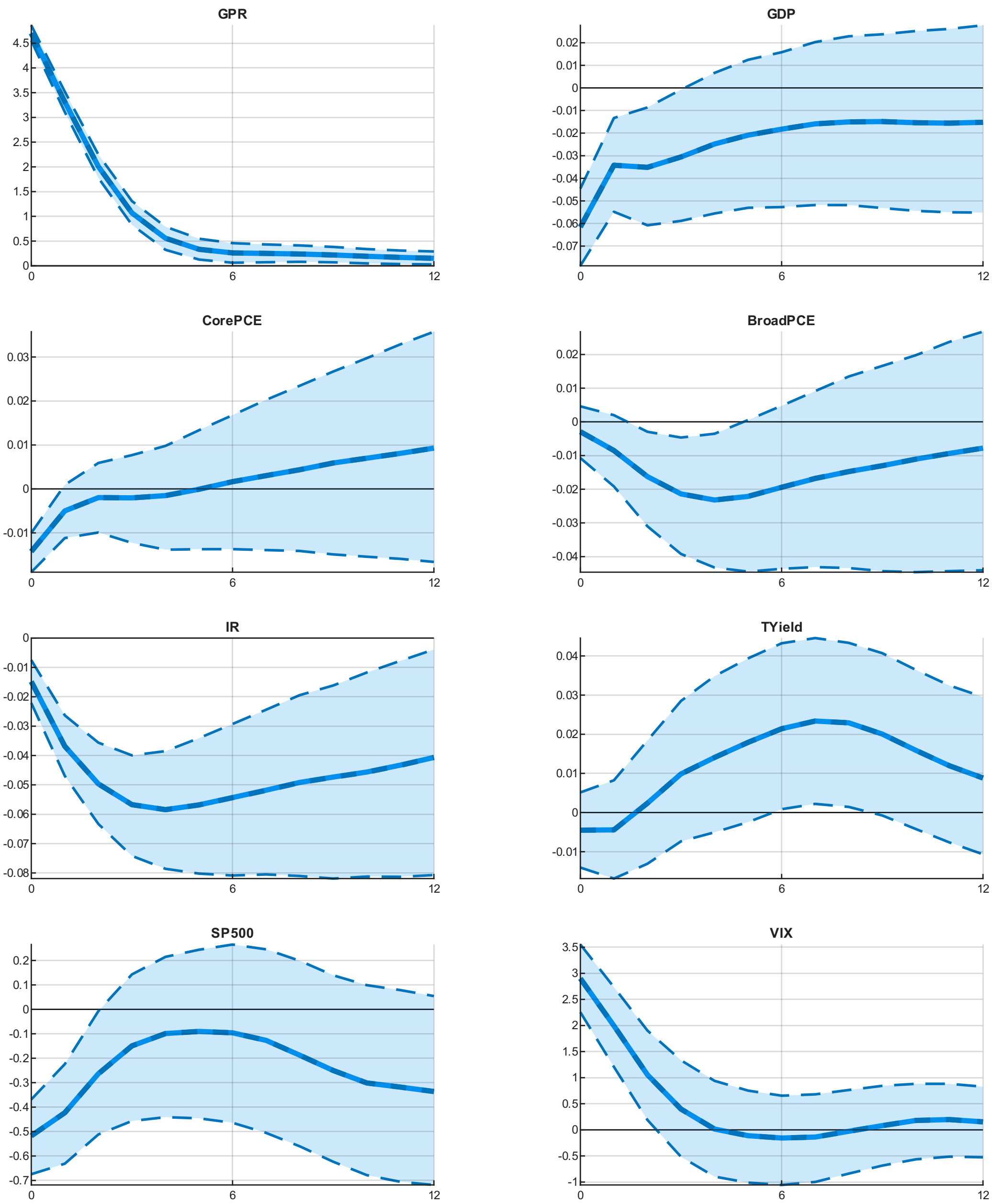


Figure 13: Coherence of the demand GPR shock evaluated on contemporaneous GDP and core PCE

Notes: Impulse responses to a one-standard-deviation innovation in the demand GPR index, identified by ordering the index first in the VAR. Solid lines and shaded areas are the posterior median and 68% highest posterior density band, computed over the full posterior. The dashed lines are the corresponding median (thick) and 68% band (thin) computed over only the coherent draws, i.e. the posterior draws whose contemporaneous responses satisfy the assigned signature: GDP and core PCE falling. GPR, GDP, core PCE, broad PCE, S&P 500, and VIX are in percent; the policy rate (IR) and 10-year yield (TYield) are in percentage points. The horizon is in months. Variable definitions and the identifying assumptions are detailed in Section 3.3; the coherence criterion is defined in Section 2.2.

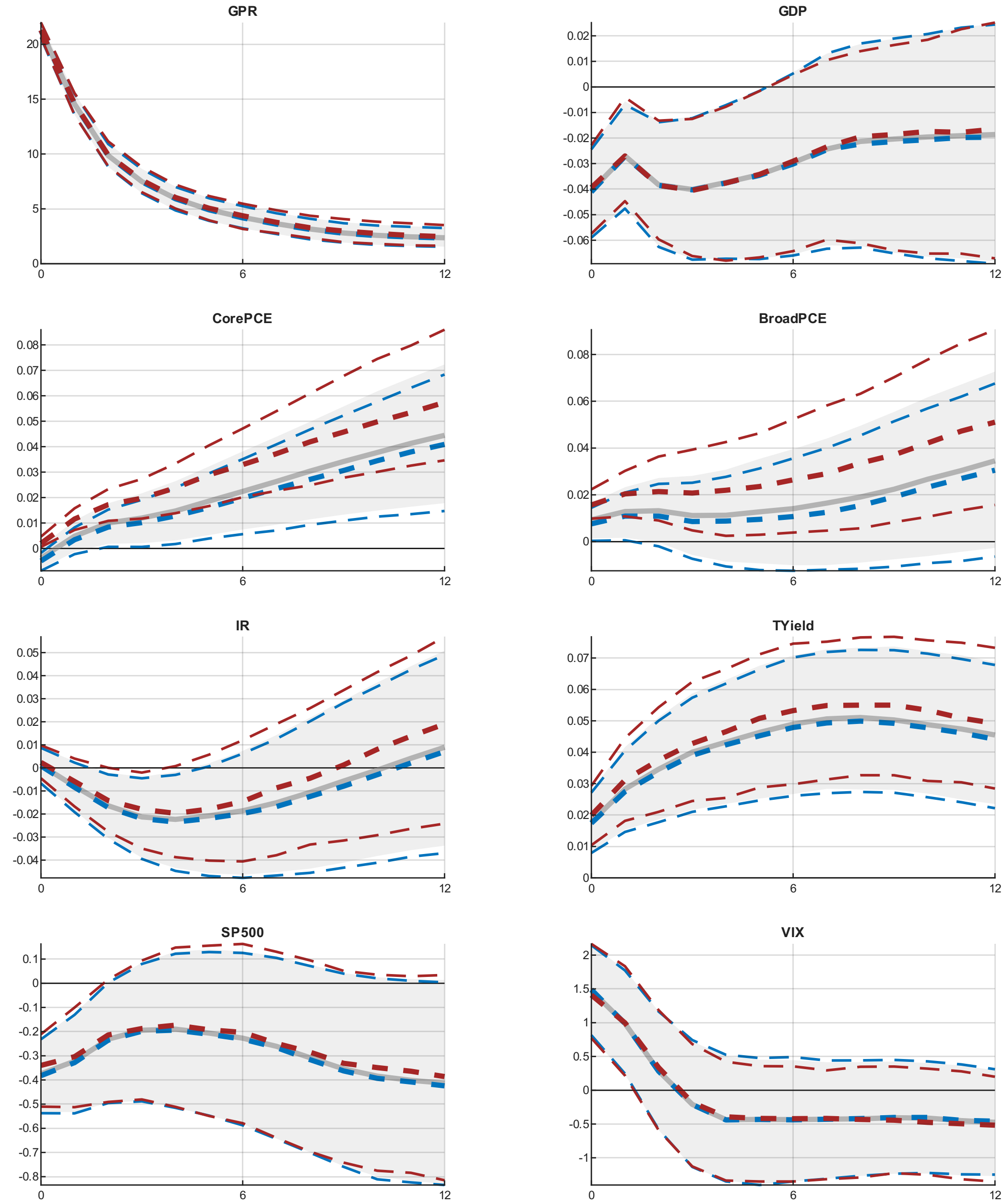


Figure 14: Coherence of the recursively identified Caldara-Iacoviello GPR shock evaluated on GDP and core PCE

Notes: Impulse responses to a one-standard-deviation innovation in the Caldara–Iacoviello GPR index, identified by a recursive (Cholesky) ordering with the index ordered first. The solid line is the posterior median and the grey shaded area the 68% highest posterior density band, computed over the full posterior. Overlaid are the responses computed over only the draws whose contemporaneous GDP and core PCE responses satisfy the supply signature (red: GDP falling, core PCE rising) and the demand signature (blue: GDP and core PCE falling together); thick dashed lines are the medians and thin dashed lines the 68% highest posterior density bands of these conditional sets. The signature is evaluated on GDP and core PCE only. GPR, GDP, core PCE, broad PCE, the S&P 500, and the VIX are in percent; the policy rate (IR) and the 10-year yield (TYield) are in percentage points. The horizon is in months. Variable definitions and the identifying assumptions are detailed in Section 3.3.

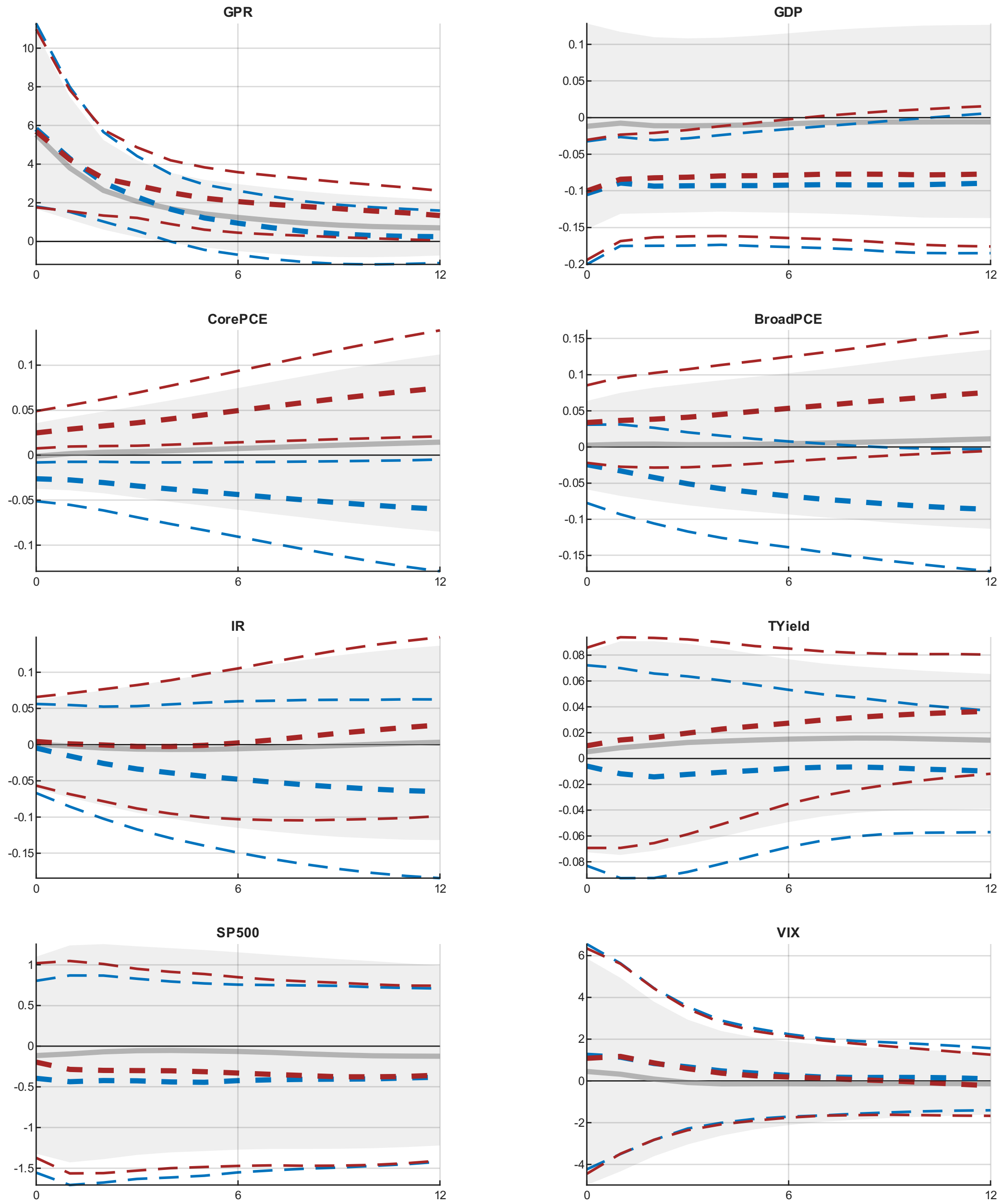


Figure 15: Sign-restriction identification on the Caldara–Iacoviello GPR index: supply and demand shocks

Notes: Impulse responses to shocks identified from the Caldara–Iacoviello GPR index by sign restrictions on the contemporaneous responses. The grey solid line and shaded area are the posterior median and 68% highest posterior density band of the base cloud—all rotations whose contemporaneous index response is positive (GPR $>$ 0), with no further restriction. The supply shock (red) additionally restricts GDP to fall and core PCE to rise; the demand shock (blue) additionally restricts GDP and core PCE to fall together. Thick dash-dotted lines are the medians and thin dashed lines the 68% highest posterior density bands of these two restricted sets. Each restricted set is the subset of the base cloud satisfying its signature: outside the restricted responses (broad PCE, the policy rate, and the 10-year yield) the red and blue bands coincide with each other and with the grey base cloud. GDP, core PCE, broad PCE, the S&P 500, and the VIX are in percent; the policy rate (IR) and the 10-year yield (TYield) are in percentage points. The horizon is in months. Variable definitions and the sign-restriction scheme are detailed in Section 3.3.