Dallari, Pietro; Gattini, Luca

Working Paper

How severe are European regulatory stress test scenarios? A probabilistic calibration for the euro area

EIB Working Papers, No. 2026/01

Provided in Cooperation with:
European Investment Bank (EIB), Luxembourg

Suggested Citation: Dallari, Pietro; Gattini, Luca (2026) : How severe are European regulatory stress test scenarios? A probabilistic calibration for the euro area, EIB Working Papers, No. 2026/01, ISBN 978-92-861-6088-2, European Investment Bank (EIB), Luxembourg, https://doi.org/10.2867/0689043

ECONOMICS – WORKING PAPERS 2026/01

HOW SEVERE ARE EUROPEAN REGULATORY STRESS TEST SCENARIOS?

A PROBABILISTIC CALIBRATION FOR THE EURO AREA

European Investment Bank

HOW SEVERE ARE EUROPEAN REGULATORY STRESS TEST SCENARIOS?

A PROBABILISTIC CALIBRATION FOR THE EURO AREA

European Investment Bank

How severe are European regulatory stress test scenarios? A probabilistic calibration for the euro area

EIB Working Paper 2026/01
January 2026

European Investment Bank
98-100, boulevard Konrad Adenauer
L-2950 Luxembourg

This is a publication of the EIB Economics Department.
economics@eib.org
www.eib.org/economics

Authors
Pietro Dallari p.dallari@eib.org, Luca Gattini l.gattini@eib.org.

About the Economics Department
The mission of the EIB Economics Department is to provide economic analyses and studies to support the Bank in its operations and in the definition of its positioning, strategy and policy. The department and its team of economists is headed by Debora Revoltella, director of economics.

Disclaimer
The views expressed in this publication are those of the authors and do not necessarily reflect the position of the European Investment Bank. EIB working papers are designed to facilitate the timely exchange of research findings. They are not subject to standard EIB copyediting or proofreading.

For further information on the EIB's activities, please consult our website, www.eib.org. You can also contact our Info Desk, info@eib.org.

Published by the European Investment Bank.

Printed on FSC® Paper.

pdf: QH-01-25-095-EN-N ISBN 978-92-861-6088-2 ISSN 2599-736X doi: 10.2867/0689043

CONTENTS

Acknowledgements

The authors thank Andrea Brasili, Jose Guedes, Carlo Pavanello, Apostolos Apostolou, Alessandra Donini, Arndt Gerrit Kund, and Martin Iseringhausen, as well as participants at the World Finance conference 2025, the 14th International Conference of the Financial Engineering and Banking Society, the ESM/SUERF/Bruegel workshop "Is Europe Prepared for Extreme Events? Risks – Resilience – Policy Responses", and seminar participants at the European Investment Bank for helpful comments. The views presented in the paper are those of the authors only and do not necessarily represent the views of the European Investment Bank.

ABSTRACT

This paper applies the Growth-at-Risk (GaR) framework to assess downside risks to euro area GDP growth, and it examines its usefulness for stress testing. Supervisory stress-test scenarios are often criticized for being either too mild or implausibly severe, raising questions about calibration. This is consequential for banks' business plans as well as for systemic financial stability. We conduct a pseudo real-time evaluation of European Banking Authority (EBA) adverse scenarios published in the last decade, establishing a probabilistic benchmark against which their scenario severity can be evaluated. We find that, except for the 2021 and 2023 rounds, EBA adverse scenarios consistently lie below the 10th percentile threshold. After the pandemic shock, scenario severity and model-implied risks have moved in opposite directions, with GaR estimates pointing to declining downside risks and EBA scenarios becoming increasingly severe. However, these still fall within the models' probability distributions and therefore represent plausible—if extreme—realizations of downside risk. Supporting exercises attribute downside risks primarily to financial stress in the short-term, while medium-term horizons are shaped by term structure and housing or credit channels more. Taken together, these results suggest that GaR can serve as a transparent, data-driven complement to expert judgment in stress-test scenario design—helping to balance severity with plausibility and enhancing scenarios' credibility for financial stability assessments.

Keywords: Growth-at-Risk (GaR); stress testing; scenario calibration; quantile regression; macro-financial conditions.

JEL codes: C22, C53, E32, E44, G21, G28.

INTRODUCTION

In the past decade, a series of major shocks have marked the end to the so-called Great Moderation—a period of reduced volatility in inflation and output growth from the mid-1980s to the mid-2000s. The 2008 global financial crisis exposed vulnerabilities in the interconnected global economy, followed by the euro area sovereign debt crisis. The COVID-19 pandemic in 2020 brought severe disruptions to supply chains and economic activity. Since 2022, the war in Ukraine has fuelled inflation and geopolitical fragmentation. Most recently, renewed uncertainty around U.S. trade policy and rising concerns over fiscal sustainability have added to the volatility of the economic outlook.

The variety of recent shocks, along with the resulting increase in uncertainty, have made it harder for policymakers to design and implement effective policy responses. With greater volatility in inflation and output growth, central banks face more difficult intertemporal trade-offs, and governments must contend with a wider distribution of fiscal risks. Financial regulators, in turn, must safeguard the stability of banks and financial institutions under conditions where macro-financial linkages are more unstable and where the distribution—not just the expectation—of macroeconomic outcomes matters for capital planning. The need to monitor systemic risks, ensure adequate capital and liquidity buffers, and maintain public confidence in the financial system has become more pressing as macrofinancial conditions have become less predictable.

In response, advances in modelling techniques have sought more robust tools for forecasting and risk assessment (Strasser et al., 2021). One line of work adapts traditional models—such as multivariate time series models and stochastic general equilibrium models—by mitigating the impact of large outliers (Lenza and Primiceri, 2022). Others incorporate high-frequency data or pandemic-specific features to capture the unique patterns observed in recent years (Cardani et al., 2021). Machine-learning techniques offer promise for non-linearities, though challenges remain around scalability, transparency, and integration into policy frameworks.

In parallel, the Growth-at-Risk (GaR) framework has gained prominence for analysing downside risks. GaR models apply risk management ideas to macroeconomic conditions (Adrian et al., 2019), estimating the distribution of future output growth and the probability of extreme negative outcomes conditional on a set of controls. A key insight is that financial conditions have asymmetric effects, particularly in downturns (IMF, 2019). This provides a natural bridge to stress testing, where the credibility and usefulness of supervisory scenarios hinge on their placement within the relevant risk distribution rather than on point expectations.

On the regulatory side, stress testing has become a central component of financial stability policies since the Global Financial Crisis. In Europe, EU-wide stress tests were introduced in 2009, and are now regularly coordinated by the European Banking Authority (EBA), in cooperation with the European Systemic Risk Board (ESRB) and the European Central Bank (ECB). These exercises assess the resilience of banks and the broader financial system under severe yet plausible macrofinancial scenarios. While primarily used for regulatory oversight, stress tests have also been a critical element in strengthening both the quality of capital and the sophistication of risk management at large banks. By forcing institutions to confront model-based loss projections under adverse conditions, and by disclosing the results, supervisors created incentives for banks to improve governance, data quality, and forward-looking risk management practices. Significantly higher risk-based capital ratios since the global financial crisis are only one dimension of this progress. The other is qualitative, with banks now better able to measure, monitor, and adjust their exposures.

While there is broad consensus that scenarios should be more severe than average expected negative events—lying in the tails of the distribution to ensure rigor, build confidence, and support the development of robust forward-looking business plans in both tranquil and turbulent times—scenario design itself remains somewhat a

debated area. Some view stress scenarios as excessively severe (Bonucchi et al. 2022), others consider them too mild to be informative (Borio et al. 2014). The credibility and policy usefulness of stress testing hinges critically on the calibration of adverse scenarios: if these are perceived by senior managers or supervisors as implausible, they risk being dismissed and lose disciplinary power. Striking the right balance requires combining multiple perspectives. Darne et al. (2024) emphasize grounding initial shocks in the economic expertise of the institution and its staff, while ensuring empirical anchoring through historical experience. We argue that a third dimension is equally essential: calibration should reflect contemporaneous macro-financial conditions at the time decisions are made. These three elements—forward-looking hypothetical narratives, historical precedent, and real-time context—are complementary. Hypothetical scenarios embed forward-looking risks that have not yet materialised but are judged plausible, some of which may eventually occur in reality—such as geopolitical escalations or abrupt policy shifts. Historical shocks provide objective anchors. Real-time calibration aligns severity with current macrofinancial vulnerabilities, enhancing credibility and informativeness.

Building on these considerations, we formalise a transparent, data-driven benchmark for scenario severity. Unlike current EBA guidance, which does not prescribe how overall severity should be assessed, we anchor scenarios in their likelihood of realisation, conditional on both historical tail events and prevailing macrofinancial conditions. This should not be interpreted as a call for publishing the full architecture of supervisory models. Rather, the value lies in reporting where an adverse scenario sits relative to the contemporaneous distribution of risks. This kind of disclosure enhances credibility by showing markets and banks how supervisory judgments map onto observable data, while avoiding the pitfalls of excessive transparency that could ossify models or invite gaming (Barr, 2025).

The paper investigates how the GaR framework can contribute to stress-test design and risk assessment, offering an empirically grounded approach to calibrate downside risks in line with prevailing economic conditions. In particular, we evaluate whether the severity of adverse macroeconomic scenarios used in EBA stress-testing exercises aligns with the empirical distribution of downside risks implied by GaR. Our analysis focuses on real GDP growth—the anchor variable in stress-test design—and conduct a pseudo real-time application (based on final data, ignoring subsequent revisions) to estimate conditional distributions at different points in time. We then benchmark the EBA adverse scenarios against the lower tail of these distributions, focusing on the 10th percentile (GaR10) as a probabilistic plausibility threshold.

Two additional components strengthen the assessment. First, we study internal validity—density calibration and distributional behaviour—by examining whether GaR10 aligns with realized lower-tail outcomes at the expected frequency and by tracking how fitted densities respond to changing macro-financial conditions. Second, we evaluate external performance by mapping GaR-implied probabilities into recession alerts and comparing classification accuracy with simple benchmarks. We implement both a fixed baseline threshold to preserve comparability and an optimal threshold (Youden-J) to probe the best attainable balance between sensitivity and specificity.

The main results preview is as follows. We trace how the severity of EBA adverse scenarios has evolved relative to prevailing macro-financial conditions and model-implied risks. GaR10, the 10th percentile of projected three-year cumulative GDP growth, varies substantially across vintages—benign in 2016 and 2018, collapsing during the COVID-19 shock, and only partially normalising thereafter. By contrast, EBA adverse scenarios remain systematically more severe than model-implied downside risks, with the notable exception of 2021 and 2023, when pandemic-related uncertainty temporarily aligned regulatory assumptions with the thickening of the lower tail. In most rounds, the probability of observing outcomes worse than the EBA adverse path is low, underscoring the consistently conservative calibration of supervisory scenarios. Following the pandemic, scenario severity and model-implied risks have moved in opposite directions, with GaR estimates pointing to declining downside risks and EBA scenarios becoming increasingly severe.

Interestingly, standard model specifications struggle to interpret the simultaneous GDP collapse and surge in financial stress as a coherent left-tail event. Smaller models, in particular, generate densities centred around positive growth, mechanically extrapolating mean reversion rather than capturing the depth of the shock. To address this limitation, we introduce a small set of parsimonious nonlinear, regime-dependent terms that allow the conditional distribution to respond more sharply in downturns and periods of elevated stress. These extensions keep the structure deliberately light but permit the left tail to deepen in a way that better mirrors pandemic dynamics.

Complementary diagnostics indicate that the divergence between model-implied risks and supervisory assumptions is not merely a matter of point estimates. Total uncertainty and downside risk are moderate in the early rounds, widen substantially in 2021, compress again in 2023—consistent with the fall-and-rebound pattern and strong policy backstops—and by 2025 return to levels comparable to 2016–2018. Recession probabilities display a similar profile: modest through 2018, elevated in 2021 and 2023, and back to single-digit levels by 2025. Distributional comparisons further show that richer model specifications deliver more credible and stable densities across regimes.

Taken together, these findings show that alignment between supervisory scenarios and model-based tail risks can vary considerably over time, with the pandemic marking a clear episode of convergence. In other rounds, EBA adverse scenarios sit deeper in the left tail than GaR10, raising proportionality questions while remaining within the empirical probability space. This underscores the value of GaR as a complementary benchmark: not as a substitute for narrative design, but as a systematic anchor ensuring that severity remains appropriately high while empirically grounded.

Beyond benchmarking supervisory scenarios, we also examine the mechanics and robustness of GaR. Few results emerge. First, drivers of downside risk are horizon-dependent: near-term forecasts are dominated by GDP and financial stress (CISS), while at medium–long horizons credit, interest rates (especially long-term), and asset prices gain prominence, with financial stress attenuating by twelve quarters. Second, GaR10 calibration is broadly accurate, with realized growth falling below the 10th percentile at close to the nominal frequency; richer specifications capture skewness and tail thickening better, at the cost of added complexity. Third, mapping GaR to CEPR-dated recessions yields high discriminatory accuracy at short horizons and moderate-to-strong performance at longer ones, with alerts that are more concentrated and better timed than simple univariate benchmarks. Robustness checks excluding the pandemic confirm these patterns: removing 2020–2024 lowers forecast errors and smooths densities, while weakening long-horizon performance mainly for parsimonious models; richer models maintain calibration more effectively.

The rest of the paper is organized as follows. Section 2 outlines the methodological framework. Section 3 describes the data. Section 4 benchmarks EBA scenario severity against GaR-implied distributions. Section 5 presents tail-risk diagnostics and robustness checks. Section 6 concludes.

1 EMPIRICAL FRAMEWORK

Growth-at-Risk (GaR) models were primarily designed to gauge the likelihood of adverse outcomes for economic activity. In essence, GaR models provide a probabilistic assessment of future output growth by fitting a continuous distribution to a set of estimated quantiles, obtained by regressing output growth on one or more explanatory variables. This differs from alternative empirical approaches that focus on point forecasts of conditional means. The bottom part of the probability distribution captures the severity of adverse outcomes. GaR models are state dependent—in the sense that the assessment of risks to growth is conditional on the information set available to the econometrician and summarized by the chosen set of explanatory variables—and non-structural, so they are not designed to recover causal links. Rather, they provide a reduced-form yet tractable framework to estimate the likelihood and severity of an economic slowdown, especially well-suited for comparative statics analysis.

The paper of Adrian et al. (2019) presents a foundational framework, making a seminal contribution to the literature. The authors aim to enhance the forecasting of future output growth by focusing on the conditional distribution of potential outcomes, emphasizing the critical role of financial conditions in identifying downturns.

The seminal work of Adrian et al. (2019) has been extended in a number of other directions. One line of research applies the GaR methodology to countries other than the United States. For example, Figueres and Jarocinski (2020), Alessandri et al. (2019) and Alessandri and Di Cesare (2021), De Lorenzo Buratta and Maia (2022), Hartigan and Wright (2021), and Adrian et al. (2022) examine the relationship between output growth and alternative financial conditions indicators for the euro area, Italy, Portugal, Australia, and a panel of advanced and emerging market economies, respectively. Generally, these studies confirm the short-term comovement of financial stress and output growth.

Other studies have explored the out-of-sample forecasting performance of GaR models compared to alternatives, such as those by Brownlees and Souza (2021) and Plagborg-Møller et al. (2020). These analyses find that quantile regressions perform similarly to GARCH models in forecasting, and that financial variables offer little additional information beyond what is already captured by other real macroeconomic indicators. Iseringhausen (2021) models output growth using a stochastic volatility specification, replacing the assumption of Gaussian shocks with a noncentral t-distribution driven by a skewness parameter that varies over time as a function of macrofinancial conditions. He shows that this approach enhances forecast accuracy against competing models in a panel of 11 OECD countries. Recent methodological advances include the work of Chavleishvili and Manganelli (2024), who introduce a quantile vector autoregression (QVAR) model, an extension of the traditional vector autoregression (VAR) framework. Using euro area data, the authors show that the QVAR model significantly outperforms standard VAR models in forecasting, particularly at the lower quantiles.

Another strand of the literature expands the analysis to include a broader set of explanatory variables, incorporating factors that capture the structural characteristics of an economy, such as trade openness, financial sector size, the public spending ratio, and government effectiveness (Gachter et al., 2023). These studies show that structural differences across countries result in varying degrees of sensitivity to changes in financial risks. Additionally, factors related to the macroprudential stance—such as capital requirements, macroprudential policy indexes, and other capital- and borrower-based measures—are also examined to explore the trade-offs between mitigating risks and fostering economic growth (Skrinjaric 2024).

Finally, Furceri et al. (2024) extend the Growth-at-Risk methodology to sovereign debt analysis by introducing a "Debt-at-Risk" (DaR) framework and proposing a new metric for sovereign surveillance, encompassing a broad information set beyond the purely mechanical debt accumulation rule. Using quantile regressions, they estimate

the distribution of future debt-to-GDP ratios conditional on prevailing macro-financial conditions, focusing on fiscal risks and systemic debt sustainability.

Related efforts within the growing literature on scenario calibration and probabilistic stress testing include Barbieri et al. (2022) and Bonucchi and Catalano (2022), who have applied time-series models to assess scenario plausibility. In contrast, our use of the GaR framework emphasizes the full distribution of risks, offering a natural link to the tail-focused logic of stress testing. By quantifying the probability associated with specific adverse outcomes, the GaR provides a rigorous foundation for assessing whether stress scenarios are appropriately severe, given the macro-financial environment in which they are applied. Recently, Adrian et al. (2025) have developed a formal methodology to synthesize judgmental scenarios with statistical forecasts, addressing a longstanding challenge in policy institutions: reconciling narrative-based scenario analysis with model-based risk assessments. Their approach leverages Bayesian predictive synthesis and entropic tilting to assign probabilistic weights to scenarios based on their concordance with a reference distribution—typically derived from GaR models. In contrast to their approach, this paper benchmarks regulatory stress scenarios against the lower tail of a GaR-implied growth distribution, using it as an external plausibility check. Both contributions share the objective of grounding macroeconomic risk assessments in probabilistic and data-consistent frameworks.

Model estimation proceeds in two main steps. We first establish the empirical relationship between output growth and the explanatory variables by estimating conditional quantile regressions. We then map the set of fitted quantiles into a smooth predictive distribution by fitting a flexible parametric family. The following two subsections describe these steps in more detail.

1.1 Quantile regressions

Regression methods explain or predict the value of an outcome variable using information in a vector of predictors. Among the most popular, ordinary least squares (OLS) targets the conditional mean. When the relationship varies across the outcome distribution or interest focuses on the tails, the mean is an inadequate summary. Quantile regressions address this by estimating conditional quantiles—for example, the median or the 10th percentile—given the same information set. As such, they offer a more complete statistical analysis of the relationship among random variables, as they estimate the relationship between a specific quantile of the outcome variable and the explanatory variables.

Let y_{t+h} denote cumulative GDP growth h-quarters ahead and let X_t collect the predictors observed at time t. For a quantile index $\tau \in (0,1)$, we estimate

$$Q_t(y_{t+h}|X_t) = \alpha_\tau + X'_t\beta_\tau, \qquad (1)$$

aligning the timing so that X_t is strictly pre-determined with respect to y_{t+h}.[1] Parameters $(\alpha_\tau, \beta_\tau)$ solve

$$(\hat{\alpha}_\tau, \hat{\beta}_\tau) = argmin_{\alpha,\beta} \sum_t \rho_\tau(y_{t+h} - \alpha - X'_t\beta), \quad \rho_\tau(u) = u(\tau - 1\{u < 0\}), \qquad (2)$$

[1] Using X_t aligns with standard GaR practice that conditions on current financial and macroeconomic conditions. Since we benchmark scenarios around EBA's cut-off dates in the past, all relevant data releases are observed.

the standard "check loss" criterion of Koenker and Bassett (1978). We estimate the model on a grid $\tau \in \{0.10, 0.15, \ldots, 0.90\}$ and at horizons $h \in \{1, 4, 12\}$, but our primary object for the EBA comparison is $h = 12$. The fitted conditional quantiles are

$$\hat{Q}_\tau(y_{t+h}|X_t) = \hat{\alpha}_\tau + X'_t\hat{\beta}_\tau. \quad (3)$$

Quantile regressions are linear in parameters within each quantile, but allow the slope vector β_τ to vary with τ, thereby capturing distributional heterogeneity and non-linear effects across the outcome distribution. In finite samples, fitted quantiles at nearby τ can cross. We mitigate this by estimating on a coarse grid (10th to 90th percentiles) and, crucially, by the second-step distributional fit, which imposes a monotone quantile function. For inference and diagnostics, we use bootstrap standard errors.

1.2 Parametric fitting

The second step converts the discrete set of fitted conditional quantiles into a smooth predictive distribution, providing a probability region which allows to quantify downside risks as the area in the left tail of the distribution. Following Adrian et al. (2019), we adopt a flexible skewed-t family with four parameters—location μ, scale σ, skewness (shape) λ, and degrees of freedom (tail thickness) ν. The skewed-t family has proven effective in macro-finance applications because it captures both fat tails and skewness, features that are empirically salient for multi-year cumulative growth.

For each t, we choose $(\mu_t, \sigma_t, \lambda_t, \nu_t)$ to minimize the squared distance between the quantile regression-implied conditional quantiles and the theoretical quantiles of the skewed-t distribution:

$$(\hat{\mu}_\tau, \hat{\sigma}_\tau, \hat{\lambda}_\tau, \hat{\nu}_\tau) = argmin_{\mu,\sigma,\lambda,\nu} \sum_{\tau \in T}[\hat{Q}_\tau(y_{t+h}|X_t) - F^{-1}(\tau; \mu, \sigma, \lambda, \nu)]]^2], \quad (4)$$

where $F^{-1}(\cdot)$ denotes the skewed-t quantile function. This delivers a time-t conditional CDF $\hat{F}(\cdot\,|X_t)$ and associated quantiles $q_\tau(X_t)$. We interpret GaR10 as $q_{0.10}(X_t)$ and the conditional median as $q_{0.50}(X_t)$. Expected-shortfall and expected-longrise are computed as integrals of $F^{-1}(\tau|X_t)$ over the relevant tail probability.

2 DATA

To investigate the sources of risk to euro-area growth, we extend the scope of existing GaR applications—often focused on the relationship between output growth and financial stress indicators (Alessandri et al., 2019; Krygier and Vasi, 2021; Buratta and Maia, 2022)—by incorporating additional channels. The predictor set combines an indicator of systemic financial stress (level), the term structure of interest rates (4-quarter level changes of short- and long-term rates), private-sector credit, house prices, equity prices, the real effective exchange rate, goods exports, and oil prices.

This choice reflects the transmission of monetary and financial conditions to real activity through balance-sheet, valuation, income, and external-competitiveness channels. In particular, the term structure of interest rates reflects the monetary policy stance; credit flows, house prices, and equity prices all contribute to economic growth and household wealth, having played key roles in previous economic cycles; the real effective exchange rate serves as a measure of the euro area external competitiveness, while goods exports and oil prices act as proxies for external demand and the commodity cycle, respectively; finally, we align with many related studies by adopting the Composite Indicator of Systemic Stress (CISS) as the preferred measure of financial stress.[2]

The dependent variable is three-year cumulative real GDP growth. All predictors are measured at time t and aligned so that they condition outcomes over $t+1, \dots, t+12$, avoiding look-ahead bias. In most exercises, variables are left in economically interpretable units (levels or growth rates), rather than standardized, to facilitate interpretation. The dataset covers the period from 1980Q1 to 2024Q4. Overall, our dataset balances a long European sample, a parsimonious regressor set, and sources of upside and downside risks.

The quarterly frequency and long span are chosen to cover multiple business cycles, improving the identification of tail behaviour and the stability of conditional quantiles. Extending the history requires carefully stitching sources where official euro-area series start later. Specifically: (i) real short-term and long-term rates are taken from the AWM in the early years and spliced to the ECB's quarterly real 3-month Euribor and real 10-year benchmark yields; (ii) equity prices use the ECB EURO STOXX 50 (quarterly average) and are back-cast to 1980 by chaining to the S&P 500 converted to euro with historical exchange rates (DEM/USD and the fixed DEM/EUR conversion); (iii) credit to private non-financial sector is taken from the ECB BSI dataset, with the history extended backward by proxying with M3 growth in the pre-1997 period;[3] (iv) house prices use the ECB RESR euro-area property price index (EA20 fixed composition); (v) exports of goods come from IMF Direction of Trade Statistics (EA to the world), converted to euro; (vi) the real effective exchange rate is the BIS Real Narrow EER for the euro area (via FRED); (vii) oil prices use WTI (converted using DEM/USD pre-1999 and USD/EUR thereafter); (viii) financial stress is the ECB quarterly CISS; (ix) real GDP is taken from Eurostat from 1995 onward and extended backward with the ECB AWM aggregate to provide a consistent long history.[4] All nominal quantities are deflated using the euro-area GDP deflator (World Bank). Table 1 lists series, sources, and start dates.

[2] The CISS is a daily indicator capturing volatility across several market segments, including bank and non-bank financial intermediaries, money markets, securities (equities and bonds) markets as well as foreign exchange markets. The CISS puts relatively more weight on situations in which stress prevails in several market segments at the same time, with the intent to better capture systemic stress. Compared to earlier works, we can use a version of the CISS that has been recently made available and goes back to 1980.

[3] There is no material difference between M3 and alternative measures of money. The correlation coefficient with credit to the private sector in the common part of the sample is 0.98 for M3 compared to 0.93 for M1. The dynamics of credit and monetary aggregates started diverging meaningfully since implementation of quantitative easing by the ECB, but were remarkably close before.

[4] The ECB AWM is an estimated structural macroeconomic model for the euro area developed for the assessment of economic conditions, forecasting and policy analysis. The model stopped being updated in 2017, but researchers can still refer to the historical databases to complement official statistics compiled by Eurostat and the European Central Bank.

3 GAR FORECASTS AND EBA ADVERSE SCENARIOS

In this section, we apply the GaR framework in an out-of-sample forecasting exercise to assess the distribution of risks to euro area GDP growth in pseudo real time.[5] We consider four alternative GaR model specifications. The first follows a baseline specification commonly found in the literature, relating future GDP growth to a parsimonious set of predictors, notably GDP and an index of financial turbulence. The others include progressively broader sets of macro-financial variables, as detailed in the data section, to capture a richer information set.[6]

Throughout the out-of-sample exercises we adopt pseudo real-time cutoffs tied to EBA stress-testing vintages. These define the information set X^{OS} at which the predictive distribution and its decomposition are evaluated. In particular, to establish a realistic policy benchmark, we select the years in which the EBA conducted its stress-tests and align our forecast horizon with EBA's three-year timeframe. As EBA does not explicitly disclose the information set underlying the exercise, we define the cutoff date for the out-of-sample forecasts as either the last quarter (Q4) or the average growth over the first three quarters (Q1-Q3) of the year preceding the respective EBA stress test. For instance, when benchmarking the 2025 scenario, we condition the forecasts on data available up to 2024Q4, or the 2024Q1–Q3 average. This approach is reasonable, as it aligns with EBA's need to incorporate economic developments throughout the year while retaining a buffer to finalize its assessment before publication.

Since 2011, the EBA has conducted several rounds of stress testing, which have progressively evolved in complexity, scope, and severity. Our analysis concentrates on those conducted in the past ten years. A central element of these exercises is the design of adverse macro-financial scenarios intended to capture a range of potential shocks—including asset price collapses, geopolitical tensions, and pandemics (Table 2). These scenarios are deliberately severe, reflecting potential low-probability tail events that could have significant implications for the financial system. This approach is conceptually aligned with the probabilistic structure of the GaR framework, which offers a technically sound benchmark for evaluating the plausibility of EBA's downside scenarios.

3.1 GaR-based assessment of scenario plausibility

Table 3 presents an overview of the evolution of downside scenario severity in successive EBA stress test rounds, benchmarked against model-implied risk distributions based on the GaR framework. The top panel reports the three-year cumulative decline in GDP under each EBA adverse scenario (AS). The remainder of the table summarizes the out-of-sample GaR projections, including measures of downside risks, total uncertainty, the probability of a recession, and the likelihood of outcomes more severe than those envisaged in EBA's adverse scenarios.

The numbers in Table 3 correspond to the most pessimistic GaR model specification, identified as the one yielding the lowest GaR10 estimate among the four models considered. This choice provides a conservative benchmark for assessing the severity of EBA adverse scenarios, while ensuring internal consistency across all associated risk metrics. The degree of heterogeneity across model specifications is examined in more detail in the next section.

[5] While not a genuine real-time analysis, as it relies on final data vintages, this detail is unlikely to materially affect our results given that the macroeconomic series we use are typically subject to only limited revisions.

[6] Specifically, the first model uses GDP and CISS; the second adds long-term rates, private sector credit and house prices; the third further includes equities and the exchange rate; the last model adds short-term rates, exports and oil prices.

A few key insights emerge. First, GaR10—the 10th percentile of the projected three-year cumulative GDP growth distribution—shows substantial variation across stress-testing rounds, reflecting shifts in macro-financial uncertainty and the business cycle. In the 2016 exercise, GaR10 is positive at 0.2%, consistent with benign post-crisis conditions, recovers to 2.5% in 2018, but drops sharply to -5.8% in the 2021 round, underscoring the severe tail risks triggered by the COVID-19 shock. In 2023 GaR10 improves to -3.4%, reflecting better but still fragile conditions, and in 2025 it stabilizes around 2.8%.

These findings are confirmed when using cutoff dates based on Q1–Q3 data instead of year-end vintages. The overall pattern remains broadly unchanged: GaR10 is 1.4% and 2.1% in the 2016 and 2018 rounds, drops to -7.3% in 2021, partially recovers to –4.0% in 2023, and reaches 4.3% in 2025. The ranking of stress episodes and turning points is preserved, and so is the conclusion that downside risks peaked during the pandemic and have since been gradually reabsorbed.

Building on GaR10, additional insights emerge when considering model-implied measures for total uncertainty and downside risks. In the early rounds, both uncertainty and downside risk are moderate. In the 2016 vintage, total uncertainty stands at 9.4 percentage points, with downside risk accounting for 4.4 points, consistent with a symmetric distribution tilted toward moderate negative outcomes. These values decline somewhat through 2018 (7.1 and 4.6), but reach 17.0 and 8.5 points in 2021, capturing the extreme dispersion of outcomes at the height of the pandemic. The 2023 round shows a moderate improvement to 12.6 and 5.5, respectively, before easing further to 4.2 and 3.1 in 2025, suggesting a gradual recentering of expectations but some persistent fragility in the lower tail.

The Q1–Q3 vintages convey, once again, a similar story. By 2018, total uncertainty and downside risk stand at 6.8 and 4.4 percentage points, respectively. In 2020 they peak at 16.8 and 8.4, then moderate to 14.0 and 6.6 in 2023, and eventually reach 2.8 and 2.1 by 2025. The sequence, therefore, reflects alternating phases of stress amplification and normalization, consistent with evolving macro-financial conditions.

Across both vintages, recession probabilities remain contained through 2018, consistent with positive GaR10 readings and moderate downside risks. They rise materially in 2021, reaching 34.0% (Q4 cut-off) and 43.2% (Q1–Q3 cutoff), reflecting the deep left-tail realizations of the pandemic. In subsequent rounds, recession probabilities moderate to 30% in the 2023 round, and drop further to 3.8% and 1.7% by end-2024, confirming that the extreme risks of the pandemic shock had been reabsorbed.

The comparison between the EBA adverse scenario (AS) and model-based GaR results offers key insights into the relative severity and plausibility of supervisory scenarios. In each EBA round, the three-year cumulative GDP contraction embedded in the AS can be compared against the GaR10. The key question is whether GaR10—which represents a lower bound to the empirical growth distribution conditional on prevailing macro-financial conditions—aligns with or significantly deviates from the EBA's stress assumption.

The results show a clear pattern: with the exception of the 2021 and 2023 rounds, the EBA adverse scenarios are consistently more severe than the corresponding GaR10 levels. In 2016 and 2018, GaR10 ranges between 0.2% and 2.5% (Q4 cut-offs), while the EBA adverse scenarios assume GDP contractions between -1.7% and -2.4%. The implied probability of growth below the EBA scenario is correspondingly low—below 5% in all cases—placing those scenarios deep into the left tail.

The 2021 round marks a clear exception. The GaR10 estimate for Q4 2020 is -5.8%, worse than the EBA adverse scenario of -3.6%, and the probability of a worse outcome rises to 16.8% (or 23.3% using the Q1–Q3 cut-off). In this case, the EBA scenario appears broadly aligned with tail risks. The model also appears to reflect the exceptional volatility of the pandemic period, with total uncertainty reaching 17 percentage points. We return on this point in the next section, when we consider alternative model specifications and cutoff dates.

In 2023, GaR10 remains negative (–4% in Q1-Q3) but retraces to milder values, the overall distribution narrows—total uncertainty declines from 17.0 to 12.6 percentage points—yet downside risks and other tail metrics remain elevated. Interestingly, the EBA adverse scenario moves in the opposite direction, becoming more severe (-5.9%). Counterfactual analysis confirms that the improvement in GaR10 between the 2021 and 2023 stress-testing rounds stems almost entirely from the normalization of macro-financial conditions rather than from changes in the estimated model coefficients.[7]

In 2025, the pendulum swings back. GaR10 turns positive (2.8% and 4.3% for the Q4 and Q1-Q3 cut-offs, respectively), while the EBA adverse scenario becomes, one more time, even more severe (-6.3%). As a result, the implied probability of breaching the AS threshold drops to less than 1%.

The implications are twofold. First, these results point to a widening gap between supervisory stress assumptions and data-driven assessments of plausibility, raising questions about the realism and proportionality of recent EBA scenarios. Second, we note that all EBA adverse scenarios remain within the models' probability space, and the likelihood of outcomes more severe than those assumed by the EBA is small but positive and broadly stable across rounds. Therefore, these probabilities are consistent with the regulatory expectation that adverse scenarios should be severe yet still lie within the range of plausible outcomes.

3.2 Distributional shifts and model robustness across stress test vintages

Figure 1 shows how model-implied distributions of three-year cumulative real GDP growth evolve across EBA stress-test vintages and cutoff dates (Q4 and Q1–Q3), highlighting differences across model specifications and their alignment with supervisory assumptions. In particular, each distribution is derived from a GaR specification of increasing complexity—from the parsimonious two-variable model based on GDP and financial stress (light blue) to progressively richer versions including macro-financial indicators such as credit, housing, equity, and external variables (darker blue). Vertical blue bars mark the model-implied GaR10 thresholds, while the orange bar denotes the severity of the corresponding EBA adverse scenario.

The 2016 and 2018 rounds exhibit compact, nearly symmetric distributions centered around mildly positive growth. The 2016 case reflects the steady post-crisis recovery, with all GaR10 levels in positive territory and the EBA scenario positioned deep in the left tail—several percentage points below any model-implied outcome, suggesting an extremely conservative calibration. By 2018, the distributions broaden modestly and shift slightly leftward, consistent with a slowdown in euro-area growth outlook and higher market volatility at the time. Yet, the EBA adverse scenario remains far more severe than the model-implied tail risks. These results confirm the conclusions in Table 3 discussed above.

The 2021 vintage is a good example of how modelling choices can shape tail behaviour when confronted with the unprecedented volatility of the COVID-19 period. The largest model, which incorporates the full range of macro-financial variables, produces a very wide density, reflecting the exceptional volatility of the pandemic period and the high dispersion in tail risks. As reported in Table 3, its GaR10 metric still captures a pronounced left-skew, signalling that the model effectively detects the surge in downside risks, despite the unusually large uncertainty. At the opposite extreme, we can see in Figure 1 that the simplest specification—relying only on GDP and financial stress—remains narrowly concentrated around positive growth, failing to register the depth of the pandemic collapse. The intermediate models behave in between: their distributions broaden somewhat but remain centered around relatively high medians, suggesting that linear relationships estimated on pre-pandemic

[7] Using 2023 betas with 2021 data yields nearly identical tail risks to those observed in 2021, whereas using 2021 betas with 2023 data reproduces most of the 2023 improvement. This points to a strong state-of-the-world effect (as captured at cutoff dates) and limited contribution from model re-estimation.

data struggle to capture the nonlinear nature of the 2020 shock. This pattern is robust to variations in the cutoff date.

In particular, in Q2 2020, euro-area GDP contracted by roughly 14% year-on-year while the CISS index spiked to 0.36, from an average of about 0.20 in the previous quarter. Yet, all models, especially the smaller ones, fail to translate this dramatic deterioration into a clear leftward shift of the conditional growth distribution. This reflects the structure of the estimated coefficients. Taking the simplest model as an example, the GDP coefficient at the 10th percentile is –1.25 and that on CISS –11.98. When these are combined with the observed Q2 2020 inputs, the linear predictor at the 10th quantile yields (ignoring the constant for simplicity):

$$\hat{y}_{0.10} = \beta_0 + \beta_{GDP}\, GDP + \beta_{CISS}\, CISS$$

or

$$(-1.25 \times -14) + (-11.98 \times 0.36) = +13.2$$

Interestingly, the large positive GDP contribution dominates the negative effect of CISS, causing the conditional left tail to move up instead of down. This counter-intuitive response stems from the estimation sample, where mild recessions dominate: the model learns that negative GDP growth tends to be followed by a rebound over the medium-term (three-year horizon), embedding an implicit mean-reversion that distorts tail behaviour in extreme downturns. In addition, the effect of CISS is highly nonlinear, but is not properly captured by the model: its pre-2020 variation is small, so a 0.16 jump beyond its historical range is treated as a linear extrapolation, rather than a systemic event. In other words, the model was simply never trained to interpret simultaneous GDP collapse and market panic as a coherent left-tail shock.

To address these limitations, we introduce a set of parsimonious nonlinear transformations, in the spirit of Lenza and Primiceri (2022), designed to capture regime-dependent dynamics. The first modification allows the conditional distribution of future growth to respond asymmetrically to expansions and contractions by interacting output growth with regime indicators that isolate downturns. This relaxes the linearity constraint implicit in standard quantile regressions, enabling the conditional distribution to exhibit asymmetric responses around the business-cycle turning points. A second extension incorporates an interaction term between financial stress and negative growth regimes, ensuring that the impact of financial turbulence is amplified when real activity is already weak—consistent with empirical evidence on state-dependent propagation of financial shocks. Finally, we define a threshold-based measure of excess stress that activates only when the market stress indicator exceeds its historical norm, thus allowing tail risk to increase more than proportionately during systemic events. Together, these extensions preserve the model's parsimony while restoring an empirically plausible degree of nonlinearity and asymmetry in the conditional growth distribution.

Formally, we augment the regressor set described in equation 3) with downturn and threshold interactions as follows:

$$Q_t(y_{t+h}|Z_t) = \alpha_\tau + W'_t\theta_\tau + X'_t\beta_\tau + \beta_{g,\tau}g_\tau + \eta_\tau(g_\tau D_t) + \beta_{s,\tau}s_\tau + \zeta_\tau(s_\tau D_t) + \lambda_\tau s^+_t \qquad (5)$$

where $Z_t \equiv (W'_t, g_\tau, g_\tau D_t, s_t, s_\tau D_t, s^+_t)$ is the augmented regressor vector; $D_t \equiv 1\{g_\tau < 0\}$ is the downturn indicator; the marginal effect in contractions is captured by $\beta_{g,\tau} + \eta_\tau$; the effect of financial stress rises to $\beta_{s,\tau} + \zeta_\tau$ when growth is negative; and λ_τ loads excess stress $s^+_t \equiv \max(0, s_\tau - c)$ with *c* set at 0.25. Everything else in the estimation process remains the same.

Figure 2 highlights how introducing some degree of regime dependence fundamentally improves the characterization of downside risk. The inclusion of state-dependent terms generates a visible thickening of the

lower tail. The GaR10 bars move leftward across all specifications, bringing model-implied risks much closer to the EBA adverse scenario. The effect is most pronounced for the largest model, whose distribution captures both the magnitude and dispersion of the pandemic-era uncertainty. However, smaller and intermediate specifications also exhibit meaningful leftward adjustments, with GaR10 values around –3 to –4 percent compared with positive values in the linear case. All models converge within a realistic range of downside risk, suggesting that the nonlinear structure corrects the mechanical—and ultimately inaccurate—prediction that characterized the linear formulations. The median of the distributions remains broadly unchanged, confirming that the adjustment operates primarily through the left tail rather than the central tendency, and that the nonlinear effects act as conditional-volatility amplifiers rather than mean shifters. The pattern is robust across cutoff definitions: using Q1–Q3 data rather than the end-of-year point produces similar left-tail behaviour.

More generally, these results also illustrate the value of richer model specifications. Incorporating additional transmission channels enhances the likelihood of identifying economic relationships whose coefficients converge to uniformly negative values in the tails, restoring monotonicity and ensuring economically coherent tail behaviour. In a policy setting, incorporating multiple specifications as a form of model averaging, can be especially important to enhance the credibility and proportionality of the assessment.

In contrast to the 2021 round—where nonlinear terms were required to capture asymmetric tail dynamics—the 2023 results, obtained from the baseline linear specification, display a visible improvement in downside sensitivity (Figure 1). The implied distributions remain narrower and more symmetric than under the nonlinear setup, yet their lower tails are better developed across all specifications. For the largest model, the GaR10 level falls to about –4½% under the Q4 cutoff and to roughly –5% under the Q1–Q3 average. The EBA adverse scenario, at around –5½%, now lies just beyond the lower tail of the model-implied distribution. The smaller and intermediate models remain more concentrated, with peaks around modestly positive growth, but they too exhibit leftward shifts relative to earlier rounds. This behaviour suggests that the inclusion of the pandemic period in the estimation window has partially taught the model the nonlinear structure of severe macro-financial stress. The overall pattern is consistent across cut-off conventions. The Q1–Q3 averages generate slightly wider densities, reflecting the smoother macro-financial environment of early 2022 relative to year-end volatility, yet the positioning of GaR10 thresholds remains stable. Compared with 2021, the 2023 results thus represent a gradual return to normalcy in both the level and shape of the risk distribution—less extreme than during the pandemic, but still characterized by elevated uncertainty and persistent tail vulnerability.

By 2025, the estimated distributions shift decisively to the right, reflecting the normalization of macro-financial conditions and the reabsorption of pandemic-era tail risks. Across all specifications, the bulk of the probability mass is now centered around positive cumulative growth, between roughly 3% and 6%, while the left tail becomes thin and symmetric. The corresponding GaR10 levels rise back into positive territory—around 1½% to 3% depending on model complexity—signalling a return to broadly benign risk conditions. This rightward realignment is visible under both the Q4 and Q1–Q3 cutoffs, suggesting that the result is not driven by timing but by a structural reversion of the data. Even the most comprehensive specification, which previously captured the fat-tailed, skewed dynamics of 2021, now exhibits a narrow bell-shaped distribution with limited dispersion. The smaller and intermediate models converge tightly around similar medians, indicating a generalized compression of macro-financial volatility in the estimation sample. Against this backdrop, the EBA adverse scenario—at roughly –6.3%—lies far into the left tail, several standard deviations below any model-implied GaR10 threshold. This configuration implies a low empirical probability of realization, even under the most pessimistic model, and represents a return to the pre-pandemic pattern where supervisory assumptions exceed model-consistent downside risks by a wide margin.

Together, these figures visually substantiate the key message emerging from the quantitative GaR metrics discussed earlier. Whereas the earlier analysis relied on a single specification to illustrate the mechanism, the results here span four distinct model types. While supervisory stress scenarios have remained severe across EBA

rounds, their alignment with model-consistent downside risks has fluctuated considerably over time. The 2021 and 2023 rounds appear as an inflection point where empirical tail risks and regulatory assumptions converged. By contrast, both earlier and later rounds feature adverse scenarios that fall well below model-implied 10th percentiles, highlighting the potential role of GaR-based benchmarks in guiding more proportionate, evidence-based scenario design.

3.3 Decomposing downside risks

The discussion so far has evaluated the plausibility of supervisory scenarios against model-implied risk distributions. A natural next step is to open the "black box" of the GaR forecast and attribute the resulting downside risk—the gap between the conditional median and the left tail—to economically meaningful drivers. Our objective in this section is precisely that: to decompose the GaR10–median gap into contributions from groups of predictors that summarize financial conditions (comprising short- and long-term interest rates, and the CISS), the asset side (private credit and house prices), external factors (including the exchange rate, exports, and oil prices), and the GDP block.

Methodologically, after estimation of the linear quantile regressions and mapping from discrete conditional quantiles to a full predictive distribution by fitting a skewed-t family at each point in time, we attribute movements in quantiles to blocks of regressors via counterfactuals. Let X^{OS} denote the out-of-sample information set at the end of the assessment year, and let X^{BASE} be a baseline vector constructed from the recent median of each regressor over a window ending at the cutoff. For each probability $\tau \in \{0.10, 0.50\}$ we compute $q_\tau(X^{BASE})$ and $q_\tau(X^{OS})$. We then form one-block-at-a-time, symmetrized counterfactuals that switch each group of predictors from the baseline to the out-of-sample value (and vice-versa), holding all other groups fixed in turn. The average of these two paths defines the contribution $c_g(\tau)$ of group g to the change in the τ-quantile between X^{BASE} and X^{OS}.[8]

The object of interest is the downside-risk gap at the out-of-sample date:

$$q_{0.10}(X^{OS}) - q_{0.50}(X^{OS}).$$

We rewrite it as:

$$q_{0.10}(X^{BASE}) - q_{0.50}(X^{BASE}) + \sum_g (c_g(0.10) - c_g(0.50)$$

Where $q_{0.10}(X^{BASE}) - q_{0.50}(X^{BASE})$ identifies the baseline gap.

The window over which the regressors' medians are estimated is set at two years, with the intention to reflect normal conditions in the current regime, not a long-run average of the entire sample. Using recent medians stabilizes the reference point and prevents the intercept from absorbing historical asymmetries that are irrelevant for present-day risks. The baseline gap should then be interpreted as the model-implied GaR10–median difference at normal conditions today. This term captures structural features that generate left-skew even in neutral states—compounding of quarterly losses, the fact that recessions are sharper than expansions, and the persistence of bad shocks, to list a few examples. The remaining blocks measure how current macro-financial conditions push the tail farther below (or closer to) the median relative to that baseline. In this way, the decomposition separates a time-varying but slowly moving background asymmetry from the state-dependent

[8] Two remarks qualify the interpretation. The contributions are conditional and contemporaneous. They explain the configuration of quantiles at the cutoff date rather than structural causal effects. The small non-additivity introduced by the skewed-t mapping is intentionally omitted, as it is quantitatively negligible and does not alter the ranking of drivers.

effect of financial, credit and housing, external, and GDP-related blocks.[9] The panels in Figure 3 illustrate the results from this decomposition.

In 2016, the baseline gap remains clearly negative (about –3¼ percentage points), confirming the intrinsic left-skew of the growth distribution even in a relatively tranquil phase of the cycle. The financial block provides the largest offset, adding roughly 1 to 1¼ points under both cutoffs and thereby narrowing downside risks. Asset prices contribute a smaller positive effect of around ¼ point, external factors exert a mild drag (–0.4 under Q1–Q3 cutoff and –0.2 under Q4 cutoff), and the GDP block is broadly neutral. Netting these contributions yields a moderate GaR10–median gap of about –2½ to –3 percentage points, smaller than in earlier rounds and consistent with reduced tail risks during the mid-cycle recovery. The Q1–Q3 and Q4 decompositions are very similar, indicating that in a relatively calm regime the precise cutoff choice matters little.

In 2018, downside risks re-emerge modestly. The baseline gap remains negative at about –2⅔ percentage points, and the sectoral decomposition shows that most blocks slightly deepen the lower tail. Financial conditions turn mildly adverse (around –0.4), consistent with the gradual normalization of global yields and the tapering of ECB asset purchases. Asset prices add a small negative contribution of roughly –0.2 percentage points, while external factors exert an additional mild drag (between –0.25 and –0.17 points across the two cutoffs). The GDP block provides a small positive offset of roughly 0.15 percentage points. Taken together, these components generate a GaR10–median gap of about –2¾ percentage points, somewhat larger than in 2016 but still moderate by historical standards. Results are stable across the Q4 and Q1–Q3 cutoffs.

In 2021, the downside tail reaches its maximum extent in the sample. The GaR10–median gap widens to roughly 5½ percentage points under the Q1–Q3 cutoff and just above 5 percentage points under Q4 cutoff, confirming the exceptional tail risks carried over from the pandemic period. The baseline gap remains large and negative (about –4 percentage points), reflecting the deep output losses accumulated during 2020. The decomposition indicates that financial conditions provided some temporary support—a positive contribution of about 1 percentage point in early 2021 and close to zero by year-end—while external and asset-price channels remained slightly negative. The GDP block continues to be the dominant source of downside risk, subtracting nearly –2¾ points under the Q1–Q3 cutoff and about –½ points under Q4 cutoff. This component captures the lingering imprint of the 2020 lockdown-induced contraction and the volatility that followed. The nonlinear terms reinforce this effect, with past weakness in activity translating into elevated conditional downside risks. In this sense, the model embeds the pandemic's legacy through the state variables observed at the cutoff.

In 2023, the growth-at-risk profile largely normalises relative to the pandemic years. The baseline gap remains negative (about –2¾ percentage points), but the decomposition shows that several blocks substantially offset this inherited weakness. The financial block provides the largest positive contribution, adding between 2.1 and 3.1 points depending on the cutoff, consistent with tighter spreads, improved market conditions, and resilient bank funding. Asset prices also make a positive contribution (around 0.6 to 1.3 points), while external factors act as a mild drag (–0.5 to –1½ points), reflecting lingering energy-related pressures and soft global trade. The GDP block is slightly negative (–0.4 to –0.6 points), signalling that activity had not fully regained pre-pandemic momentum at the cutoff date. Netting these effects yields a small GaR10–median gap of roughly –0.8 percentage points under the Q1–Q3 cutoff and –0.6 under Q4 cutoff, indicating that by 2023 the distribution of expected growth had returned to near-symmetric conditions, with downside risks modest and substantially smaller than in earlier rounds.

[9] It is useful to note that our decomposition has a clear linear counterpart. In the quantile-linear approximation, the GaR10–median gap equals the intercept difference plus the regressor values times the difference in quantile slopes, $\Delta\beta = \beta_{0.10} - \beta_{0.50}$. Algebraic recentering at X^{BASE} simply moves $\Delta\beta' X^{BASE}$ into the baseline gap and expresses each group's influence as $\Delta\beta'_g (X^{OS} - X^{BASE})$. Our implementation retains this intuition while ensuring distributional coherence across quantiles.

In 2025, the baseline gap is only mildly negative (about –0.6 percentage points), indicating that the unconditional distribution remains close to symmetric. However, the state-dependent drivers at the cutoff point deliver a deterioration. The financial block provides the largest negative contribution—between –1.6 and –2.7 points across the two cutoff dates—reflecting tighter monetary conditions, elevated real rates, and a repricing of funding conditions relative to the 2023–24 median values. The asset block also weighs on the lower tail (around –1½ to –2¼ points), consistent with softer housing dynamics and more volatile or subdued equity valuations. External factors range from neutral to mildly negative, depending on the cutoff, while the GDP block offers a modest positive offset of roughly +0.2 to +0.4 points, in line with resilient activity and stabilising labour markets. Netting these forces yields a GaR10–median gap of –1 point under the Q1–Q3 cutoff and –2.6 points under the Q4 cutoff, underscoring that by the 2025 vintage the balance of risks has tilted back to the downside, driven primarily by tighter financial conditions and asset-market headwinds.

Viewed in aggregate, the decomposition offers a clearer economic anatomy of downside risk across markedly different environments. A persistent structural asymmetry in multi-year growth provides the backdrop, but the state-dependent channels—financial conditions, asset-market dynamics, external environment, and the evolving cyclical position—govern how far the lower tail departs from that baseline. The comparative evidence across years underscores that the identity and strength of these channels rotate with macro-financial regimes, and that cutoff sensitivity becomes consequential only in periods of pronounced turbulence. This systematic mapping from observables to tail outcomes gives economic content to the probabilistic benchmarks developed earlier and sharpens the interpretation of supervisory scenario severity. It also delineates the levers—most notably financial conditions and credit-housing channels—through which policy interventions are likely to exert the greatest influence on left-tail exposure in future stress-test rounds.

4 EXPLORING MODEL STRUCTURE AND BEHAVIOUR

This section takes a step back from the main scenario-based assessment to investigate how GaR models behave across different dimensions. The goal here is to explore the behaviour of the model—its sensitivity to explanatory variables, forecast horizon, and sample selection—and to gauge how these affect its capacity to capture downside risks.

Unlike traditional forecast evaluation, there is no single "ground truth" for scoring predictive densities. The aim therefore is not to minimize forecast errors in a narrow sense, but to judge if and how alternative models map the evolving distribution of macro-financial risks. To that end, the analysis in the following subsections is guided by three complementary questions: (i) do GaR estimates vary systematically with underlying macro-financial conditions, (ii) do the predicted densities display features—such as skewness, kurtosis, and tail thickness—that align with economic intuition during periods of stress, and (iii) can shifts in the risk distribution be meaningfully traced back to observable drivers?

This diagnostic perspective is valuable not only for guiding model development, but also to enhance transparency in policy applications, where choices about model specification, variable inclusion, and forecast horizon can materially shape the perceived severity of risks.

4.1 Multivariate quantile regressions: patterns across models and horizons

We begin by estimating multivariate quantile regressions using successively richer sets of explanatory variables.[10, 11] Since we consider a total of 10 potential regressors, we start from the smallest model and add one variable to the previous set in a nested structure. For each model, we compute coefficient estimates at the 10th and 90th percentiles of the growth distribution, as well as standard OLS estimates, across 1-, 4-, and 12-quarter ahead horizons. Table 4 reports both the point estimates and their associated p-values, facilitating comparisons of sign, magnitude, and significance across specifications. For clarity and readability, Table 4 shows only results from a selected subset of these models—specifically, number 3, 5, 7, and 10.

This is complemented by Figure 4, which displays the full distribution of estimated coefficients for each variable, quantile, and forecast horizon. The figure combines uncertainty from the model (across the 10 specifications) and parameter estimation (via bootstrapping), allowing for a richer representation of the behaviour of each regressor across the growth distribution.

Several patterns emerge. The financial stress index (CISS) consistently exhibits large and significantly negative coefficients across all models and horizons, with the largest effects observed at the 10th percentile. At the short-term horizon (1–4 quarters), a one-unit increase in CISS is associated with a GDP contraction typically around 8

[10] For brevity, we do not report the results of univariate quantile regressions in which output growth is regressed individually on each explanatory variable. These remain available from the authors upon request. The findings are broadly consistent with those from the multivariate specifications.

[11] Quantile regression coefficients do not lend themselves to causal interpretation or cross-model comparisons, particularly when multiple regressors interact. We use them as a reduced-form, flexible description of the conditional relationship between explanatory variables and output growth across different quantiles.

percentage points in the lower tail, while the upper-tail effects are much smaller.[12] This asymmetry underlines the non-linear role of financial stress in amplifying downside risks during crisis episodes. The effect persists even after controlling for equity prices, interest rates, and credit, confirming CISS's central role as a tail-risk driver.

Private credit growth and house prices are positively associated with GDP growth at shorter horizons. These effects are strongest in the lower quantiles, albeit heterogeneous across models, implying that credit and asset valuations play a stabilizing role when growth is weak. However, their influence fades—and even reverses—at longer horizons. At the twelve-quarter horizon, both variables turn negative or insignificant, with particularly adverse effects at the lower tail. This shift aligns with the view that credit and housing booms may sow vulnerabilities that materialize only in the medium term.

Interest rates show pronounced horizon-dependent asymmetry. Short-term rates have only modest effects at one quarter, but by four quarters their coefficients are clearly negative and significant in the lower tail—consistent with policy transmission lags whereby tightening dampens demand after a few quarters. At twelve quarters, the effect of short-term rates becomes weaker and occasionally turns positive. This pattern is consistent with yield-curve dynamics whereby short rates partly capture expectations of policy reversion (tightening today often followed by easing later) with growth subsequently recovering. Long-term rates, by contrast, grow in importance with horizon: at four quarters they are statistically insignificant or even positive, consistent with yields reflecting near-term growth and inflation expectations, but by twelve quarters they turn strongly negative—particularly in the 10th percentile, where coefficients fall close to or below –1 in several specifications—reflecting the drag of persistently restrictive financing conditions on medium-run growth. In Figure 4, this pattern is visible in the downward tilt of the long-term rate boxplots across horizons.

For the Euro Stoxx 50 index, estimated effects are very small at all horizons. At one and four quarters coefficients are slightly positive but rarely significant, while at twelve quarters they turn weakly negative in some specifications. Overall, equity valuations provide at best modest near-term signals for growth and show little systematic predictive power at longer horizons.

GDP growth is an immediate and powerful near-term predictor: at one quarter, coefficients are large (around 0.4-0.6 across quantiles) and highly significant, with a slightly stronger effect in the left tail. By four quarters, the relationship flips and weakens—coefficients turn negative (−0.4 to −0.2)—consistent with mean reversion. At twelve quarters, effects are mixed: the lower-tail effect is positive and significant, while OLS and the upper tail are small and insignificant, suggesting limited average and upper-tail predictability but some mitigation of downside risk at longer horizons.

Across specifications, REER, exports, and oil prices generally have small effects. REER coefficients hover near zero and are not significant at any horizon; the lower-tail tilt at Q4 is directionally consistent with competitiveness pressures but not statistically robust. Exports are negligible at Q1–Q4, but at Q12 the lower-tail effect is sizeable and significant, indicating that weak external demand materially amplifies downside risks over the medium run. Oil price coefficients remain close to zero and statistically indistinguishable from zero across horizons; the mild negatives at Q4 do not reach conventional thresholds.

Four anchors hold: i) GDP is among the strongest near-term predictors, ii) financial stress is the dominant tail-risk driver, with the largest effects at short-to-intermediate horizons, iii) short-term rates are clearly contractionary by one year, and iv) long-term rates dominate medium-run downside risk.

To facilitate a cross-variable comparison, we compute the median regression coefficients from the bootstrap replications and normalize them by each variable's standard deviation, so values indicate deviations from the variable-specific mean (z-scores) abstracting from units and scale. Figure 5 displays these standardized estimates

[12] Note that the CISS ranges between 0 and 1, and it peaked at 0.94 during the Great Financial Crisis on December 8th, 2008. Therefore, a unit increase corresponds to an extremely severe outcome.

as heatmaps across horizons and quantiles. Lighter colours (yellow) indicate more positive effects relative to a variable's own distribution, while darker shades (dark red) denote negative deviations, or below-average effects. The plots pool all multivariate specifications, thus capturing both model and parameter uncertainty.

At the one-quarter horizon, GDP growth normalized coefficients are around +2 standard deviations across most quantiles, signalling strong near-term persistence. CISS is already below its mean in the lower tail, consistent with stress amplifying downside risks on impact. EURIBOR is modestly below its mean at the lower quantiles, while exports tilt above their mean in the lower tail. Credit and house prices are mildly above their means in weak states; REER sits close to its average influence; oil is only weakly negative.

By the four-quarter horizon, GDP growth drops below its mean across quantiles, consistent with the mean reversion identified in Table 4. The CISS retains strong negative values in the lower tail (between –1.5 and –2.0), reinforcing its role as a short-to-intermediate-horizon tail-risk driver. Credit and house prices maintain positive effects of about 0.5 to 1.0 at the 10^{th} and 25^{th} percentiles. Short-term interest rates turn more contractionary, with coefficients close to –1.0 in the lower quantiles, consistent with the lagged transmission of monetary policy. Long-term yields move above their mean, particularly in the lower quantiles, reflecting their information content about improving expectations rather than restrictive financing conditions. Exports and the REER lean below average, while oil prices remain a modest drag with a non-monotonic pattern. This intermediate horizon thus reflects a more diversified set of downside risk drivers, with macro-financial linkages playing an increasing role.

At the twelve-quarter horizon, patterns polarize across quantiles. Long-term rates display the sharpest split—bright at low and very dark at high quantiles—indicating relatively stronger lower-tail effects than their own average but pronounced headwinds toward the upper tail. In levels, these coefficients are negative overall at this horizon, consistent with restrictive financing conditions in Table 4. Short-term rates similarly flip above their mean in the lower tail and below in the upper tail, in line with policy reversion at long horizons. Credit growth, house prices, and equity prices also display negative deviations of –1.0 to –2.0 in the lower tail, pointing to the accumulation of financial vulnerabilities over time as early-cycle supports turn into sources of instability. The CISS effect attenuates compared with shorter horizons—its normalized coefficients are closer to zero or mildly negative—but level estimates remain below zero, suggesting mean reversion of stress rather than a reversal. GDP turns moderately positive (around 0.5), no longer dominates the distribution. Oil prices exhibit an asymmetric pattern, with negative effects in the lower tail and mild positives in the upper quantiles, while exports and the REER remain weakly negative.

Overall, three patterns emerge from the set of results discussed in this section. First, the predictive content of GDP is frontloaded and diminishes rapidly beyond the first few quarters. Second, financial stress (CISS) exerts a consistently negative influence across horizons, especially in the lower tail, though its peak impact is at intermediate horizons. Third, the role of macro-financial variables—especially interest rates, credit, and asset prices—intensifies at longer horizons, shaping the distribution of growth outcomes in ways that are asymmetric and horizon-dependent: short rates turn contractionary by one year, long rates dominate at three years, and early-cycle supports (credit and house prices) fade or reverse. Together, these findings support using a broad and time-varying set of predictors in GaR models to capture evolving downside risks with greater plausibility and granularity.

4.2 Quantile fit and tail risk representation

This subsection addresses two complementary aspects of model validation: first, whether the estimated GaR10 aligns with realized adverse outcomes at the expected frequency; and second, whether the predicted distributions' shape, dispersion, and asymmetries evolve in ways consistent with macro-financial conditions, particularly during systemic shocks.

To further assess the model's ability to represent downside risks, and to evaluate how well the estimated lower quantiles reflect actual tail outcomes in the data, we compare the fitted GaR10 forecasts to realized year-on-year GDP growth at the one-, four-, and twelve-quarter horizons. This is done in-sample, using model coefficients estimated over the full sample period, which allows us to assess the internal coherence and tail calibration properties of each specification.[13]

The analysis focuses on two benchmark specifications that span the range of models considered in the paper. The first is a minimalist baseline model, including only GDP growth and financial stress (CISS), in line with Adrian et al. (2019) and similar to those used in many early applications of GaR. The second is the full model, incorporating a broad set of macro-financial variables designed to capture alternative transmission channels beyond pure financial stress. These two models represent opposite ends of the spectrum in terms of parsimony and informational richness. While intermediate specifications are not reported here, one can reasonably expect their performance to fall between these two extremes.

The results indicate that the 10th percentile forecasts are broadly well-calibrated across all horizons (Table 5). At the one-quarter horizon, the share of observations where realized GDP falls below GaR10 is 9.7% for the small model and 11.4% for the large model—both close to the nominal 10% level. Calibration remains acceptable at the four-quarter horizon, with hit rates of 9.9% (small) and 8.1% (large), and at the twelve-quarter horizon with 10.4% (small) and 8.5% (large). These figures suggest that the estimated lower quantiles capture the empirical frequency of adverse outcomes reasonably well across time and horizon.

Forecast errors—defined as the difference between realized growth and GaR10—are systematically positive, as expected given the lower-tail focus of the metric. These errors increase with the forecast horizon, reflecting the natural widening of uncertainty and model dispersion. At the one-quarter horizon, mean forecast errors are 0.93 (small model) and 0.79 (large model), with median errors of 0.86 and 0.71, and RMSEs of 2.02 and 1.89, respectively. At four quarters, mean errors rise to 1.50 and 2.09, with median errors of 1.43 and 2.12, and RMSEs of 2.84 and 3.17. By twelve quarters, the underprediction becomes more pronounced, with mean errors of 1.75 (small) and 2.00 (large), median errors of 2.02 and 1.73, and RMSEs of 2.99 and 3.26. This progression confirms that GaR10 acts as a lower bound that becomes increasingly conservative with horizon, consistent with its role in identifying conditional tail risks.

Importantly, the comparison of small versus large models shows no single dominance. The large specification achieves a marginally better fit at the very short horizon (lower RMSE and smaller errors at one quarter) but tends to accumulate somewhat larger errors and higher RMSEs at longer horizons. The small model, while less flexible, produces more stable performance across horizons. Intuitively, while richer macro-financial information can enhance responsiveness in the near term, it can also introduce higher noise over longer horizons.

To complement the quantitative diagnostics on forecast accuracy and tail calibration, Figure 6 visualizes the historical evolution of conditional GDP growth densities generated by the small and large models across different forecast horizons. The figure offers a three-dimensional view of the fitted probability density function (PDF) for each quarter—providing an intuitive representation of the models' distributional behaviour over time. These visualizations help assess not only the central tendency of forecasts, but also how their shape, dispersion, and skewness evolve in response to changing macro-financial conditions.

Few insights emerge. Both models respond to systemic shocks with broader and more left-skewed distributions, particularly at the one- and four-quarter horizons. These shifts are clearly visible during the early 1980s, the Global Financial Crisis, and the COVID-19 recession, underscoring both models' ability to register increases in tail risk during periods of macroeconomic stress. The small model produces relatively smooth and compact

[13] While out-of-sample evaluation would require re-estimation at each point in time, the in-sample fit already provides meaningful insights into the plausibility and statistical consistency of the estimated risk distributions.

distributions, which helps lower RMSEs at medium and long horizons but also limits flexibility. Even in major downturns, forecast densities widen only modestly, with little thickening of the left tail. At the twelve-quarter horizon in particular, densities remain excessively symmetric and stable, showing minimal variation in dispersion or skewness and thus likely understating the probability of extreme outcomes or regime shifts. By contrast, the large model generates distributions with greater variability, sharper skewness, and more pronounced shifts in response to shocks. This reflects the broader set of predictors, which enhances responsiveness to changing conditions but also raises average forecast errors at long horizons and generates more volatile signals in tranquil periods.

Overall, while the small model delivers smoother, more parsimonious estimates with stable forecast dispersion, it tends to understate large downside risks—especially at longer horizons. This highlights the tailoring model choice to the macro-financial environment and the task at hand.

4.3 Recession forecast evaluation

Having examined the internal mechanics and distributional properties of the GaR framework, we now turn to its ability to predict macroeconomic downturns. While GaR is designed to estimate conditional growth distributions, its lower-tail estimates can be interpreted as a probability of recession, linking continuous forecasts to the binary signals used in business-cycle dating. This provides a useful diagnostic for two reasons. First, the CEPR Euro Area Business Cycle Dating Committee's chronology offers a widely accepted reference for evaluating crisis-detection models. Second, translating GaR outputs into recession probabilities and discrete "alerts" allows for a direct comparison with more traditional early-warning systems and rule-based indicators, helping to situate the framework's practical value in policy settings.

Two thresholds are used to translate probabilities into binary signals. The first is a fixed baseline cut-off ($\tau = 0.30$), chosen to be sufficiently high to limit spurious alarms while maintaining cross-model and cross-horizon comparability. The second is an optimal threshold τ^*, determined separately for each model–horizon pair by maximising Youden's J statistic.[14] This dual-threshold approach allows us to contrast the properties of an operationally stable rule with those of a statistically optimal but potentially less stable one, following the practice in related forecast evaluation work (see Chauvet and Piger, 2008).

Performance is assessed through a combination of metrics. The hit rate (true positive rate) and false alarm rate (false positive rate) measure the frequency of correct and incorrect recession signals under each threshold. The area under the ROC curve (AUC) provides a horizon- and threshold-independent summary of the model's ability to rank recession and expansion periods correctly, while the Brier score captures the mean squared probability error.[15] To capture operational relevance, we also compute lead-time measures, defined as the number of quarters between the last correct recession signal and the actual onset of a downturn, within a fixed pre-recession window of six quarters.

[14] Youden's J statistic provides a rule for selecting the classification threshold that jointly maximises sensitivity (the share of recessions correctly identified) and specificity (the share of expansions correctly recognised). It identifies the cut-off that offers the best overall balance between missed crises and false alarms, without imposing explicit assumptions about the relative costs of these two types of error. The statistic ranges from 0 (no discriminatory power) to 1 (perfect classification).

[15] AUC corresponds to the probability that the model assigns a higher predicted probability of recession to a randomly drawn recession quarter than to a randomly drawn non-recession quarter, thus summarizing overall ranking ability across thresholds. Values of AUC below 0.5 may appear counterintuitive, as they imply the model performs worse than random classification. In practice, this often reflects small sample sizes, where a few mis-rankings can dominate the metric, or the instability of benchmark signals whose timing does not align well with CEPR-dated recession onsets. Importantly, an AUC below 0.5 does not mean the underlying variable carries no information (a simple inversion of the signal would yield an AUC above 0.5) but highlights poor calibration. The Brier score measures the mean squared error between predicted probabilities and actual binary outcomes. It simultaneously captures calibration (whether predicted probabilities match observed frequencies) and resolution (whether predicted probabilities distinguish effectively between recession and non-recession periods), with lower values indicating better probability forecasts.

We use as benchmarks two non-parametric models that translate observable macro-financial indicators into pseudo-probabilities of recession. The first is a GDP momentum benchmark, which computes a backward-looking moving average of real GDP growth over the relevant forecast horizon. Lower recent growth translates monotonically into higher risk through a rank transformation rescaled to the unit interval. The second is a financial stress benchmark based on the CISS index, which is mapped into probabilities by ranking higher stress levels as indicating greater recession risk. These procedures deliberately avoid estimation or functional-form assumptions: they provide transparent, easily replicable heuristics grounded in two intuitive channels—real activity momentum and financial stress transmission.

The results (Table 6) show that at the one-quarter horizon, both GaR specifications deliver strong discriminatory accuracy. AUCs are 0.87 (small) and 0.83 (large), and Brier scores are low at 0.10, signalling well-calibrated probabilities. At the baseline threshold, hit rates are 0.52 for both models, with very low false alarm rates (0.05). At the optimized τ^*, hit rates improve to 0.72 (small) and 0.80 (large), while false alarm rates increase modestly to 0.06 and 0.19. Benchmarks also perform well at this horizon, in line with findings that simple momentum or stress indicators can capture near-term turning points effectively (Hasenzagl et al., 2020). Univariate GDP achieves an AUC of 0.84 and CISS 0.81, though their error profiles are less balanced. GDP has a high Brier score of 0.25 and produces very high false alarms (0.66 at baseline), while its hit rate under τ^* remains strong at 0.84 with fewer false alarms (0.24). The CISS benchmark attains very high hit rates at baseline (0.97) but this collapses to 0.66 under τ^*, and although false alarms fall from 0.65 to 0.15, signals remain excessively frequent, as seen in its high always-on share (0.60 at baseline) and long run lengths (up to 4.2 quarters). Overall, GaR achieves a better balance of accuracy and parsimony, while the benchmarks rely on nearly continuous signalling to maintain sensitivity.

At the four-quarter horizon, performance gaps widen. AUCs for GaR models decline to 0.71 (small) and 0.68 (large), with Brier scores of 0.11, but still remain above the univariate GDP benchmark (AUC 0.47, Brier 0.34) and close to the CISS benchmark (AUC 0.72 and Brier 0.28). Under $\tau = 0.30$, GaR hit rates fall to 0.27 (small) and 0.23 (large), though false alarms remain very low (0.07–0.08). Optimising τ^* restores hit rates to 0.64 (small) and 0.73 (large), but with higher false alarm rates (0.26 and 0.38, respectively). The pattern whereby raising sensitivity to capture more recessions comes with a proportional rise in false signals is well documented in the early-warning literature.[16] The GDP benchmark loses much of its classification ability at this horizon, with AUC below 0.50, high false alarm rates (0.69 baseline, 0.85 at τ^*), and poor calibration (Brier above 0.3). By contrast, the CISS benchmark retains stronger discriminatory ability (AUC 0.72) with perfect hit rates of 0.96 even under $\tau = 0.30$, but this comes with very high false alarms (0.65 baseline, 0.59 at τ^*), near-continuous signals (always-on share 0.60), and extended run lengths (up to 4.2 quarters). GaR models thus outperform GDP by a clear margin, while offering cleaner and more concentrated signals than CISS.

At the twelve-quarter horizon, the contrast between GaR specifications becomes striking. The large model achieves the highest AUC (0.79) and the lowest Brier score (0.09), while the small model's AUC drops to 0.61 with a Brier of 0.10, barely above chance. At baseline τ, both models record very low hit rates (0.11), with the small model producing no false alarms (0.00) and the large model a modest 0.06. At τ^*, hit rates rise sharply to 0.84 (small) and 0.89 (large), but false alarms also increase substantially, to 0.53 and 0.36, respectively. This illustrates the long-horizon trade-off: sensitivity can only be increased by tolerating much noisier signals. The GDP benchmark fares poorly, with an AUC of 0.43, hit rate of 0.68 at baseline falling to 0.63 at τ^*, and consistently high false alarm rates (0.70 baseline, 0.54 at τ^*). The CISS benchmark deteriorates further, with an AUC of only 0.51 and hit rates dropping from 0.74 (baseline) to 0.26 (τ^*). While its false alarms decline (0.70 to 0.07), this improvement is mechanical and comes at the cost of missing most recessions. Coverage statistics highlight GaR's advantage: at τ^*, both small and large GaR achieve coverage of 0.80 and 0.60 of the pre-recession window,

[16] For a discussion of the trade-off between early detection and false signals, see Berge and Jordà (2011) and related literature.

respectively, compared with just 0.20–0.40 for GDP and CISS. Moreover, GaR alerts remain more contained, with shorter run lengths and lower mean shares of the window, while benchmarks—especially CISS—often verge on being permanently "on."

A consistent theme in Table 6 is the difference in threshold behaviour between the GaR and the benchmarks.[17] For GaR, the optimal τ^* is typically below the baseline cutoff (0.30), reflecting the smoother probability distributions: lowering the threshold captures more recessions without an immediate explosion in false-alarm rates, thereby maximising Youden's J. The GDP and CISS benchmarks often yield τ^* values above 0.30, especially at short horizons, because their predicted recession probabilities are highly polarised—close to one during recessions and near zero during expansions. Raising the threshold in such cases helps suppress false positives without materially reducing true positives.[18]

Coverage ratios and lead times depend as much on the threshold choice as on the underlying model. At the fixed threshold ($\tau = 0.30$), GaR models issue almost no warnings at Q1 or Q12, reflecting their conservative calibration, while the GDP and CISS benchmarks cover nearly the entire pre-recession window. Optimising τ^* boosts GaR coverage substantially, raising it to 0.20–0.40 at Q1 and to 0.60–0.80 at Q12, bringing them into line with or above the benchmarks. The trade-off is that the benchmarks' coverage actually falls at τ^*, since their highly polarised signals require stricter thresholds to avoid excessive false alarms. Lead-time statistics reveal the other side of this trade-off: GDP and CISS benchmarks deliver earlier first signals—often 5–6 quarters ahead—while GaR warnings arrive closer to the recession onset, especially at long horizons where median leads remain near zero. From an operational perspective, this suggests that GaR signals are more concentrated and timed to the brink of recessions, whereas benchmarks flag downturns much earlier but at the cost of frequent false alarms. Run-length and share measures reinforce this distinction: GaR alerts are short-lived, while benchmarks, particularly CISS, tend to produce extended or near-continuous sequences of alarms, undermining their precision as decision tools.

From a policy perspective, several points stand out. First, while simple rules based on GDP momentum or financial stress can rival GaR at very short horizons, they do so by issuing frequent or continuous alerts, limiting their operational usefulness. GaR, by contrast, produces more precise and parsimonious signals. Second, at longer horizons GaR—especially the large specification—outperforms, demonstrating the value of incorporating broader macro-financial information to detect slow-moving vulnerabilities. Third, the choice of threshold strongly shapes the signal profile: fixed cut-offs deliver conservative but sparse warnings, while optimised thresholds expand coverage at the cost of noisier signals. Policymakers therefore face a trade-off between reliability and sensitivity, but GaR offers a structured probabilistic framework to manage this trade-off, clearly outperforming naïve benchmarks and providing richer diagnostics for timing, intensity, and persistence of recession risks. For stress testing exercises traditionally entailing at least a three years horizon GaR frameworks with large, yet parsimonious, specifications seem to be a relevant tool to design and assess the severity of scenarios.

[17] The table does not report the optimal thresholds τ^* directly, but their effects can be inferred from the changes in hit and false alarm rates when moving from the fixed $\tau = 0.30$ to τ^*. Explicit τ^* values are available in the replication codes, which can be provided by the authors upon request.

[18] Flatter, noisier probability series tend to push τ^* down, while sharper, more bimodal distributions tend to push it upward, with direct implications for the timing and frequency of model-issued recession alerts. This is a direct consequence of ROC geometry: for a given ranking performance, the τ^* that maximizes Youden's J corresponds to the point on the ROC curve where the slope equals the odds ratio of event to non-event frequencies. When forecast probabilities cluster near 0 and 1, the ROC curve tends to be convex with a steep initial rise, shifting τ^* upward. When probabilities are more diffuse, the curve flattens and τ^* moves toward lower values.

4.4 Pre-pandemic sample

To test robustness, we re-estimate all models on a sample ending in 2019. The unprecedented nature of the COVID-19 shock, coupled with extraordinary policy responses, could distort some historical co-movements among macro-financial variables. For instance, activity collapsed due to mandated shutdowns and then rebounded unusually fast once restrictions were lifted—dynamics that typical recessions do not display. In a standard downturn one would also observe pronounced and persistent swings in financial markets, asset prices, labour indicators, and inflation. During the pandemic, decisive policies muted or reshaped several of these channels, potentially skewing historical relationships.[19]

As in our earlier exercises, we compare quantile-regression coefficients (Appendix Table 1 versus Table 4), the distribution of these coefficients across model variants and bootstrap replications (Appendix Figure 1 versus Figure 4), and normalized median heatmaps that summarize the sign and relative strength across horizons and quantiles (Appendix Figure 2 versus Figure 5). We also inspect the time profile of conditional growth densities (Appendix Figure 3 versus Figure 6).

The broad picture is stable across samples and confirms the core set of downside-risk predictors. GDP remains a powerful near-term predictor at one quarter. In the pre-pandemic sample, its profile is steeper into the left tail and more persistent across models. Once 2021–2024 data are added, this gradient flattens and, in places, modestly weakens at the lower quantile, consistent with the difficulty of fitting smooth conditional quantiles through the sharp contraction-rebound episode. For example, at Q12 10th percentile, the coefficient falls from about 0.91 (pre-pandemic, regression #10) to 0.54 (full sample). Financial stress retains a large, negative association with low-tail growth at short to medium horizons in both samples. Pre-pandemic, 10th percentile effects are large and precisely estimated—e.g., Q1 is −7.41 (regression #7) and Q4 is −11.41 (regression #7), −10.30 (regression #10)—underscoring strong downside amplification in crises. At Q12, CISS is still sizeable pre-pandemic (−4.78, regression #7) but weakens materially when the pandemic years are included (around −1.5 and statistically insignificant). Other financial indicators behave consistently across samples. Short-term rates show the clearest lower-tail contraction at Q4, in line with policy lags. By Q12, short rates tend to flip positive in the left tail (e.g., 0.67 in regression #5 pre-pandemic), an effect that strengthens in the full sample (1.08), consistent with expectations of eventual policy reversal. Long-term rates exhibit the opposite horizon profile: relatively muted at short horizons but strongly negative at Q12 pre-pandemic (about −0.85 in regressions #5 and #10; −0.59 in #7), and even more negative with the pandemic years included (−1.33 in regression #10), highlighting a sharper term-structure signal in the full sample. Credit and house prices support growth at short horizons (e.g., Q1: credit 0.23 in regression #5; house prices 0.10 in regression #7), but their Q12 coefficients tend to turn negative or insignificant (e.g., credit −0.33 in regression #5 and −0.39 in regression #10 pre-pandemic), a pattern that persists with the pandemic included—consistent with boom-bust dynamics that materialize over the medium run. Equities (Euro Stoxx 50) are modestly positive at short horizons (0.03), becoming negative by Q12. External variables remain secondary: the REER is near zero or insignificant throughout; exports are negligible at short horizons but adverse at Q12 (e.g., −0.17 in regression #10 pre-pandemic, similar magnitudes in the full sample); oil prices are close to zero.

Overall, the pre-pandemic sample reallocates intensity rather than overturn the hierarchy of predictors. Including 2021–2024 i) flattens the cross-quantile GDP pattern (especially at Q4), ii) dilutes long-horizon CISS effects, and iii) strengthens the term-structure channel (short rates more positive at Q12, long rates more negative), while

[19] Consumption initially fell due to restrictions, only to surge as pent-up demand was released. Financial markets, which typically decline in recessions, rebounded swiftly due to massive policy intervention. Housing prices, rather than falling as they usually would in a downturn, rose in many countries due to low interest rates and shifting preferences. Labor markets also behaved atypically. Finally, inflation, which normally weakens in recessions, saw a delayed but sharp resurgence due to supply constraints and stimulus-fueled demand.

leaving intact the core ordering: CISS and the yield curve dominate downside risk at short/medium and medium/long horizons, with GDP highly predictive in the very near term and credit/housing shifting from short-run support to medium-run vulnerabilities.

The historical conditional densities reinforce these comparisons. Across horizons, the large model consistently displays more vertical texture—shorter-run ripples in density height—because it absorbs rapid swings in financial variables. This extra sensitivity helps long-horizon alignment in the pre-pandemic sample, but it also amplifies short-horizon volatility when pandemic years are included. The small model produces smoother, more concentrated surfaces, especially at twelve-quarters. Excluding the pandemic therefore tightens the entire surface and improves RMSEs markedly: densities are smoother at short horizons and more compressed at twelve quarters, particularly for the small model. The structure of the conditional distributions, however, is stable across samples. What the pandemic years add is short-horizon spikiness and medium-horizon asymmetry (left-then-right bulges), which push up RMSE and slightly soften discrimination in the short term. In the medium-term, shapes remain smooth in both vintages, with the small model most compressed and the large model more diffuse but better calibrated pre-pandemic.

Using GaR10 as a summary of downside risk, root-mean-squared errors fall markedly when the pandemic years are excluded. For the small model, RMSE declines from 2.02 to 1.06 at one quarter, from 2.84 to 2.06 at four quarters, and from 2.99 to 2.29 at twelve quarters (Table 5 versus Appendix Table 2). For the large model, the corresponding declines are from 1.89 to 0.95, from 3.17 to 2.29, and from 3.26 to 2.22. Hit rates move in the expected direction—toward the nominal 10 percent—when the pandemic is included. For example, the large model's hit rate rises from 7.4 to 11.4 percent at one quarter and from 6.7 to 8.5 percent at twelve quarters, while the small model rises from 9.1 to 9.7 percent at one quarter and from 8.5 to 10.4 percent at twelve quarters (Appendix Table 2 versus Table 5). In other words, the COVID-19 episode creates additional tail realizations that tighten alignment with the lower quantile, but it does so at the cost of higher average forecast error because the conditional quantiles must traverse a short-lived, high-amplitude disturbance.

Discrete recession-prediction results confirm the trade-off. At the one-quarter horizon, discrimination is strong in the pre-pandemic cut, with AUCs of 0.91 for both the small and large GaR models, compared with 0.87 and 0.83 once the pandemic years are included. Brier scores are essentially the same pre-pandemic (0.09) and in the full sample (0.10). Baseline hit rates for the GaR models are 0.52, unchanged across sample vintages, and well below the high values of the univariate benchmarks (0.96). Tuning the threshold to Youden's J raises hit rates to 0.83 (small) and 0.87 (large), in line with or better than the benchmarks at the cost of higher false alarms, but broadly in line with what we observe in the full sample. Median lead times at τ^* are 3.5 quarters for both models, with window coverage of 0.5—stronger than the full sample. Overall, short-horizon classification is robust across vintages, and the small GaR performs marginally better than in the full sample.

At four quarters ahead, the picture is similar. In the pre-pandemic sample AUCs are 0.77 (small) and 0.70 (large), slightly better than the 0.71 and 0.68 of the full sample. Brier scores are unchanged at 0.11 in both vintages. Baseline false-alarm rates are low (0.08 and 0.05 pre-pandemic, versus 0.08 and 0.07 full sample). Threshold tuning lifts hit rates to 0.70 (small) and 0.60 (large), broadly comparable to the full-sample levels, but also raises false alarms to 0.25 and 0.24, respectively. Baseline median leads are 3.5/3 quarters, stretching to 5 quarters at τ^* in either vintage. Coverage windows are a touch stronger pre-pandemic: 0.50 baseline and 0.75 at τ^* (versus 0.40 and 0.60/0.80 in the full sample).

At twelve quarters, the two vintages diverge more clearly. Pre-pandemic, the large GaR is especially strong (AUC 0.91), while the small GaR is weak (AUC 0.43). Brier scores are similar across vintages—0.10 (small) and 0.08 (large) pre-pandemic, versus 0.10 and 0.09 in the full sample—and baseline hit rates remain low at such long horizons. Under τ^*, the small GaR attains a hit rate of 0.29 with a 0.10 false-alarm rate, whereas the large GaR reaches 1.00 hits with 0.09 false alarms, i.e., near-perfect classification with little persistence. Median leads at

τ* stay around 3.5–6 quarters, and coverage remains high (0.50–0.75). The small model's post-pandemic gains come via a much steeper precision-recall trade-off—it improves meaningfully (AUC 0.61 and hit rate 0.84) but only by accepting a large increase in noise (false alarms 0.53)—whereas the large model softens (AUC 0.79, hit rate 0.89, false alarms 0.36 in the full sample) but retains a more balanced long-horizon discrimination..

Also in the pre-pandemic sample, the univariate GDP and CISS benchmarks attain comparatively high hit rates mainly because they signal more persistently than the GaR models, not because they discriminate more sharply. GDP always-on share is near zero (occasionally 0.25), yet it still produces longer runs and higher mean shares than GaR around downturns. CISS is stickier at the baseline threshold, with elevated but bounded false-alarm ratios (never exceeding about 0.70) and longer runs across horizons, which support hit rates but reflect a more persistent signal than GaR. By contrast, the GaR specifications keep low mean shares and short runs at baseline while maintaining competitive AUC. A caveat is that CISS under optimized τ performs well at short horizons in this pre-pandemic cut: it does so by moving to a higher-sensitivity point on the ROC curve— accepting more persistence and a higher false-alarm ratio to gain hits. Overall, the univariate rules buy hit rates by spending more time in signal, while GaR achieves similar or stronger discrimination with a leaner, less persistent alert profile.

5 CONCLUSIONS

Growth-at-Risk (GaR) provides a practical, transparent complement to supervisory stress testing by anchoring scenario severity to an empirical distribution of future GDP growth conditional on contemporaneous macro-financial information. Applied to the euro area, it allows us to benchmark European Banking Authority (EBA) adverse scenarios against model-implied tails, trace how downside-risk drivers evolve across regimes, and assess both the internal calibration and the recession-forecasting performance of the framework.

Set against an empirical benchmark of recent macro data, EBA adverse scenarios are generally more severe than GaR-implied 10th percentiles. Consistently, the model-implied probability of growth falling below the EBA adverse assumption is typically small—below 5% in all rounds except during the pandemic, when supervisory severity and empirical tail risk broadly aligned. Early- and late-vintage EBA scenarios sit deep in the left tail, yet they remain within the models' probability space. The introduction of parsimonious nonlinear, regime-dependent terms proves essential for capturing the extreme dynamics of the pandemic period, restoring a realistic degree of asymmetry in the conditional distribution.

Tail-risk diagnostics help clarify why this alignment moves over time. Total uncertainty and downside risks are moderate in the early rounds —around 9 and 4½ percentage points in 2016, easing slightly by 2018—and then widen sharply at the height of the pandemic, reaching roughly 17 and 8½ points in 2021. They moderate in 2023 (about 12½ and 5½) and decline further by 2025 (just above 4 and 3), indicating a gradual recentering of expectations but lingering fragility in the lower tail. Recession probabilities trace the same arc: contained through 2018, rising materially in 2021 (above 30%), then easing somewhat in 2023 and eventually falling to low single-digits by end-2024. Decompositions show that the locus of vulnerability rotates with regimes: post-crisis financial tightness dominates the earlier rounds; pandemic and energy-linked dislocations dominate the 2021 spike; and by 2025 the drivers tilt back toward financial conditions and asset-market channels, with external factors playing a secondary role. This mapping highlights where policy levers carry the most traction—financial conditions and housing/credit channels—in influencing downside risk.

Model validation supports using GaR as an operational benchmark. Lower-quantile calibration is close to nominal across horizons, and forecast errors widen with horizon as expected. Excluding the pandemic years reduces RMSEs—consistent with the idea that the pandemic shock inflated short-horizon dispersion while improving tail calibration by adding genuine left-tail realizations. In recession classification, GaR signals are strong at short horizons, moderate at one year, and favour the richer specification at three years. Compared with GDP- or CISS-based benchmarks, GaR produces more selective, better-timed alerts: benchmarks can rival GaR near term, but they tend to achieve higher sensitivity via more persistent signalling and higher false-alarm shares. By contrast, GaR attains similar or stronger discrimination with shorter runs and lower mean shares.

A unifying message from the decompositions and the cross-model regressions is that financial conditions are the dominant and most state-dependent driver of left-tail outcomes, with other channels rotate in importance across horizons. The CISS amplifies downside risk most powerfully at short-to-intermediate horizons and tends to attenuate by twelve quarters. The monetary stance bites with lags: short-term rates turn clearly contractionary by about one year ahead, and long-term financing conditions become the single strongest drag at three years, shaping medium-run tail risks. Credit and housing support growth when conditions are weak but increasingly flip sign at longer horizons, consistent with vulnerability build-up and balance-sheet effects. Equity prices provide modest near-term cushioning with little persistence. External factors (REER, exports, oil) are generally second-order, but can be episodically important—as in 2021 when energy and trade dislocations added to the left tail—whereas GDP momentum dominates only at short horizons and fades quickly.

Policy takeaways follow directly. Stress scenarios should remain severe to preserve systemic and single institutions financial stability, but their plausibility ought to be judged against a data-driven distribution that reflects current macro-financial conditions. In this light, most EBA adverse scenarios are markedly conservative, with 2021 and 2023 notable exceptions. Embedding GaR in scenario governance offers a transparent plausibility check and a way to diagnose which channels are driving the tail at any point in time. This can help enhancing credibility, supporting financial stability and financial institutions business plans' viability. Because model behaviour is horizon- and regime-dependent, robustness is improved by averaging parsimonious and richer specifications and by cross-checking alternative information sets—especially in turbulent periods when cut-off timing matter.

REFERENCES

Adrian, T., Boyarchenko, N., and Giannone, D., (2019). Vulnerable Growth, *American Economic Review,* 109(4), 1263-1289.

Adrian, T., Grinberg, F., Malik, S., Yu, J., (2022). The term structure of Growth-at-Risk, *American Economic Journal: Macroeconomics*, 14(3), 283-323.

Adrian, T., Giannone, D., Luciani, M., West, M., (2025). Scenario synthesis and macroeconomic risk, IMF working paper No. 25/105.

Alessandri, P., Del Vecchio, L., and Miglietta, A., (2019). Financial conditions and growth at risk in Italy, Banca d'Italia Temi di Discussione No. 1242.

Alessandri, P., and Di Cesare, A., (2021). Growth-at-Risk in Italy during the Covid-19 pandemic, Banca d'Italia Notes on Financial Stability and Supervision No. 24.

Barbieri, P.N., Lusignani, G., Prosperi, L., and Zicchino, L., (2022). Model-based approach for scenario design: stress-test severity and banks' resiliency, *Quantitative Finance*, 22(10), 1927-1954.

Barr, M.S., (2025). Preserving the Dynamism and Credibility of Stress Testing, Peterson Institute for International Economics, Washington DC.

Berge, T.J., & Jordà, Ò., (2011). Evaluating the classification of economic activity into recessions and expansions. *American Economic Journal: Macroeconomics*, 3(2), 246–277.

Bonucchi, M., and Catalano, M., (2022). How severe are the EBA macroeconomic scenarios for the Italian economy? A joint probability approach, *Journal of International Economics and Finance*, 129(4).

Borio, C., (2014). The financial cycle and macroeconomics: What have we learnt?, *Journal of Banking and Finance*, 45, 182-198.

Borio C., Drehmann M., and Tsatsaronis K., (2014). Stress-testing macro stress testing: Does it live up to expectations?, *Journal of Financial Stability*, 12(June), 3-15.

Brownlees, C., and Souza, A.B.M., (2021). Backtesting global Growth-at-Risk, *Journal of Monetary Economics*, 118(1), 312-330.

Buratta, I., Maia, M.F.D., (2022). How bad can financial crisis be? A GDP tail risk assessment for Portugal, Banco de Portugal WP No. 04.

Burnside, C. (1998). Determinants of economic growth: a cross-country empirical study, World Bank Economic Review, 12(3), 375-396.

Cardani, R., Croitorov, O., Giovannini, M., Pfeiffer, P., Ratto, M., and Vogel, L., (2021). The Euro area's pandemic recession: a DSGE-based interpretation, European Commission Discussion Paper No. 153.

Chauvet, M., & Piger, J., (2008). A comparison of the real-time performance of business cycle dating methods. *Journal of Business & Economic Statistics*, 26(1), 42–49.

Chavleishvili, S., and Manganelli, S., (2024). Forecasting and stress testing with quantile vector autoregression, *Journal of Applied Econometrics*, 39(1), 66-85.

De Lorenzo Buratta, I., and Maia, M.F.D., (2022). How bad can financial crisis be? A GDP tail risk assessment for Portugal, Banco de Portugal Working Paper No. 04

Darné O., Levy-Rueff G., and Pop A., (2024). The calibration of initial shocks in bank stress test scenarios: An outlier detection based approach, *Economic Modelling*, 136(2)

Diebold, F. X. (2007). *Elements of Forecasting* (4th edition).

Figueres, J.M., and Jarocinski, M., (2020). Vulnerable growth in the euro area: Measuring financial conditions, Economic Letters, 191(2), 109-126.

Furceri, D., Giannone, D., Kisat, F., Lam, W.L., Li, H., (2025). Debt-at-Risk, IMF working paper No. 25/86.

Gabriel, F., Henry, J., and Mestre, R., (2001). An Area-wide Model for the euro area, ECB working paper No. 42.

Gachter, M., Geiger, M., Hasler, E., (2023). On the structural determinants of Growth-at-Risk, *International Journal of Central Banking*, 19(2), 251-293.

Hartigan, L., and Wright, M., (2021). Financial conditions and downside risk to economic activity in Australia, Reserve Bank of Australia Research Discussion Paper No. 03.

Hasenzagl, T., Plagborg-Møller, M., Reichlin, L., and Ricco, G., (2020). When is growth at risk?, *Brookings Papers on Economic Activity*, Spring, 167-229.

International Monetary Fund, (2019). Growth at Risk: Concept and application in IMF country surveillance, IMF Working Paper No. 19/36.

Iseringhausen, M., (2021). A time-varying skewness model for Growth-at-Risk, *International Journal of Forecasting*, 40(1), 229-246.

Jordà, Ò., Schularick, M., and Taylor, A.M., (2015). Leveraged bubbles, *Journal of Monetary Economics*, 76(S), 1-20.

Koenker, R., and Bassett, G., (1978). Regression quantiles, *Econometrica*, 46(1), 33.

Krygier, D., Vasi, T., (2021), Macrofinancial conditions, financial stability and economic growth in Sweden – evaluating the Growth-at-Risk framework, Sveriges Riksbank Staff Memo.

Lenza, M., and Primiceri, G., (2022). How to estimate a vector autoregression after the pandemic, *Journal of Applied Econometrics*, 37(4), 688-699.

Reinhart, C.M., and Rogoff, K.S., (2009). This Time is Different: Eight Centuries of Financial Folly, *Princeton University Press*. Available at http://dx.doi.org/10.2307/j.ctvcm4gqx

Romer, C.D., and Romer, D.H., (2000). Federal Reserve information and the behaviour of interest rates, *American Economic Review*, 90(3), 429-457.

Schmidt, L., and Zhu, Y., (2016). Quantile Spacings: A Simple Method for the Joint Estimation of Multiple Quantiles Without Crossing. Available at SSRN: https://ssrn.com/abstract=2220901 or http://dx.doi.org/10.2139/ssrn.2220901

Škrinjarić, T., (2024). Growth-at-risk for macroprudential policy stance assessment: a survey, Bank of England Staff Working Paper No. 1075.

Strasser, G., Pérez-Quirós, G., Rünstler, G., and Bobeica, E., (2021). After floods and pandemics: How to obtain a meaningful forecast, VoxEU.org, 31 October.

ANNEXES

Table 1. Data used in the analysis			
Series	**Source**	**Available since**	**Notes**
Gross domestic product	Eurostat; ECB AWM (pre-1995)	Q1:1995 (Eurostat) / Q1:1980 (AWM)	AWM used prior to 1995 to extend history
Exports of goods	IMF Direction of Trade Statistics	Q1:1980	Converted to EUR; deflated with GDP deflator
Real effective exchange rate	BIS (Real Narrow EER, via FRED)	Q1:1980	Area-wide narrow REER (2020=100)
Interest rate, short-term	ECB AWM (pre-1999); ECB 3-month Euribor (real)	Q1:1999 (ECB) / Q1:1980 (AWM)	Spliced across vintages; enters as 4-quarter change
Interest rate, long-term	ECB AWM (pre-1995); ECB 10-year benchmark (real)	Q1:1999 (ECB) / Q1:1980 (AWM)	Spliced across vintages; enters as 4-quarter change
Credit to private non-financial sector	ECB BSI; ECB M3 (proxy pre-1997	Q1:1997 (BSI) / Q1:1980 (M3 proxy)	Spliced across vintages
Euro STOXX 50 equity index	ECB; back-cast with S&P 500	Q1:1986 (ECB) / Q1:1980 (S&P500)	Pre-1986 history from S&P500 with conversion in DEM/USD and USD/EUR
Composite Indicator of Systemic Stress (CISS)	ECB	Q1:1980	Quarterly frequency
House price index	ECB	Q1:1980	---
Oil price, WTI	FRED	Q1:1980	USD series, converted via DEM/USD and USD/EUR
Notes: Some series have early entries imputed back using proxies that are close substitutes and show similar trend and cyclical patterns through the common part of the sample.			

Table 2. EBA stress-testing rounds

2016	-1.7%	Reversal of risk premia
2018	-2.4%	Repricing of risk premia
2021	-3.6%	Re-emerging waves of infections
2023	-5.9%	Geopolitical polarization
2025	-6.3%	Geopolitical tensions

Notes: Severity refers to the cumulative drop in GDP over a three-year period.

Table 3. EBA adverse scenarios and risk measures across stress test rounds

	EBA Stress Testing rounds				
3-year cumulative real GDP growth:	2016	2018	2021	2023	2025
EBA adverse scenario (AS)	-1.7	-2.4	-3.6	-5.9	-6.3
	Results - Most Pessimistic Model				
Cutoff date	**Q4 2015**	**Q4 2017**	**Q4 2020**	**Q4 2022**	**Q4 2024**
GaR 10%	0.2	2.5	-5.8	-3.4	2.8
Total uncertainty	9.4	7.1	17.0	12.6	4.2
Downside risks	4.4	4.6	8.5	5.5	3.1
Likelihood of recession	9.0	3.6	34.0	30.0	3.8
Likelihood of growth below EBA AS	3.3	1.4	16.8	4.8	0.9
Cutoff date	**Q1-Q3 2015**	**Q1-Q3 2017**	**Q1-Q3 2020**	**Q1-Q3 2022**	**Q1-Q3 2024**
GaR 10%	1.4	2.1	-7.3	-4.0	4.3
Total uncertainty	8.1	6.8	16.8	14.0	2.8
Downside risks	4.2	4.4	8.4	6.6	2.1
Likelihood of recession	4.8	4.4	43.2	29.8	1.7
Likelihood of growth below EBA AS	1.8	1.7	23.3	6.0	0.4

Notes: EBA AS are those reported in the European Systemic Risk Board reports. GaR10 is a metric that quantifies the worst potential cumulative GDP growth over a specific period (e.g., 3 years in this case) that is expected to occur with a 10% probability, based on a given probability distribution of growth outcomes. Total uncertainty is the difference between the 90th and the 10th percentile of the growth distribution. Downside risks is the difference between the 50th and the 10th percentile of the growth distribution. Likelihood of recession indicates the probability that 3-year cumulative real GDP growth is negative. Finally, the bottom row shows the probability to observe growth below the EBA AS. The cutoff date is set at the last quarter (Q4) or the average values during Q1-Q3 of the year preceding the EBA stress test exercise. The numbers shown correspond to the most pessimistic model specification, i.e. the one yielding the lowest GaR10 estimate among the four models considered.

Table 4. Quantile regressions coefficients across alternative models and forecast horizons

1-quarter ahead

	Model 3			Model 5			Model 7			Model 10		
Explanatory variable	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**
REER	0.00	0.04	-0.03	0.05	0.01	0.00	0.00	-0.01	-0.01	0.01	-0.01	0.02
	0.94	0.29	0.40	0.10	0.83	0.93	0.99	0.70	0.84	0.68	0.83	0.65
Short-term rate	0.14	0.30	-0.12	-0.06	0.33	-0.24	-0.16	0.03	-0.29	-0.25	-0.13	-0.24
	0.40	0.20	0.61	0.73	0.12	0.23	0.28	0.88	0.13	0.07	0.40	0.12
Long-term rate	-0.45	-0.54	-0.22	-0.19	-0.55	0.10	0.02	0.22	0.20	0.33	0.23	0.23
	0.03	0.05	0.42	0.33	0.03	0.67	0.93	0.35	0.39	0.05	0.23	0.24
Private sector credit				0.20	0.24	0.22	0.10	0.21	-0.01	0.05	0.06	0.00
				0.00	0.00	0.00	0.09	0.01	0.87	0.33	0.32	0.98
Euro Stoxx 50				0.04	0.05	0.01	0.02	0.01	0.02	0.01	0.01	0.02
				0.00	0.00	0.22	0.03	0.46	0.13	0.27	0.16	0.03
House prices							0.14	0.08	0.20	0.08	0.07	0.15
							0.01	0.23	0.00	0.09	0.21	0.01
CISS							-5.23	-8.27	-2.58	-3.35	-4.67	-0.98
							0.00	0.00	0.10	0.00	0.00	0.45
Exports										0.05	0.01	0.02
										0.11	0.89	0.65
Oil price (WTI)										0.00	0.00	0.00
										0.62	0.43	0.71
GDP										0.41	0.56	0.38
										0.00	0.00	0.00

4-quarters ahead												
	Model #3			Model #5			Model #7			Model #10		
Explanatory variable	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**
REER	-0.01	-0.06	0.01	0.02	0.03	0.03	-0.01	-0.07	0.04	0.01	-0.04	0.02
	0.68	0.20	0.73	0.60	0.53	0.46	0.72	0.17	0.31	0.89	0.51	0.61
Short-term rate	-0.25	-0.61	-0.01	-0.39	-0.92	-0.09	-0.46	-0.51	-0.32	-0.48	-0.53	-0.39
	0.16	0.02	0.97	0.03	0.00	0.68	0.01	0.04	0.15	0.01	0.04	0.08
Long-term rate	0.34	0.79	-0.15	0.50	0.83	0.10	0.65	0.45	0.47	0.56	0.60	0.42
	0.11	0.01	0.59	0.02	0.00	0.69	0.00	0.14	0.08	0.01	0.07	0.14
Private sector credit				0.13	0.00	0.27	0.05	-0.13	0.15	0.10	-0.06	0.05
				0.03	1.00	0.00	0.48	0.22	0.10	0.15	0.59	0.61
Euro Stoxx 50				0.03	0.05	0.00	0.01	0.02	0.01	0.03	0.03	0.02
				0.00	0.00	0.78	0.19	0.32	0.43	0.03	0.11	0.10
House prices							0.12	0.04	0.14	0.16	0.06	0.24
							0.06	0.66	0.06	0.01	0.48	0.00
CISS							-3.08	-9.82	-1.44	-4.05	-8.48	-3.11
							0.03	0.00	0.42	0.01	0.00	0.09
Exports										0.04	0.09	0.00
										0.35	0.14	1.00
Oil price (WTI)										-0.01	-0.02	-0.01
										0.18	0.12	0.45
GDP										-0.38	-0.33	-0.22
										0.00	0.06	0.15

12-quarters ahead												
	Model #3			Model #5			Model #7			Model #10		
Explanatory variable	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**
REER	0.03	-0.04	0.02	0.02	-0.03	0.02	-0.01	0.01	0.02	-0.03	-0.02	0.02
	0.41	0.40	0.57	0.48	0.57	0.51	0.72	0.78	0.21	0.43	0.73	0.39
Short-term rate	0.23	0.32	0.42	0.40	1.08	0.35	0.46	0.95	0.36	0.51	0.71	0.59
	0.27	0.24	0.12	0.06	0.00	0.20	0.03	0.00	0.19	0.02	0.03	0.05
Long-term rate	-0.16	-0.26	-0.12	-0.34	-0.82	-0.43	-0.37	-0.97	-0.44	-0.41	-1.33	-0.47
	0.47	0.36	0.68	0.13	0.01	0.14	0.10	0.00	0.14	0.09	0.00	0.14
Private sector credit				-0.20	-0.40	-0.23	-0.11	-0.19	-0.17	-0.14	-0.27	-0.18
				0.00	0.00	0.01	0.14	0.08	0.10	0.08	0.02	0.09
Euro Stoxx 50				-0.01	-0.03	0.00	-0.03	-0.04	0.00	-0.03	-0.04	0.00
				0.43	0.04	0.95	0.03	0.03	0.99	0.03	0.05	0.84
House prices							-0.14	-0.19	0.02	-0.16	-0.22	0.01
							0.03	0.04	0.77	0.02	0.02	0.95
CISS							-4.02	-1.55	-3.31	-3.67	-1.86	-2.65
							0.01	0.46	0.10	0.02	0.42	0.22
Exports										-0.05	-0.20	0.00
										0.30	0.00	0.94
Oil price (WTI)										0.00	0.00	0.01
										0.58	0.97	0.47
GDP										0.18	0.54	0.14
										0.18	0.00	0.41

Notes: Note: Coefficients and p-values are based on OLS and quantile regressions of growth distributions at the 10th and 90the percentiles. The estimation period is Q1:1980-Q4:2024.

Table 5. Model fit statistics for GaR10 forecasts at different horizons

Horizon	Hit rate		Mean error		Median error		RMSE	
	Small	Large	Small	Large	Small	Large	Small	Large
1-quarter ahead	9.7%	11.4%	0.93	0.79	0.86	0.71	2.02	1.89
4-quarter ahead	9.9%	8.1%	1.50	2.09	1.43	2.12	2.84	3.17
12-quarter ahead	10.4%	8.5%	1.75	2.00	2.02	1.73	2.99	3.26

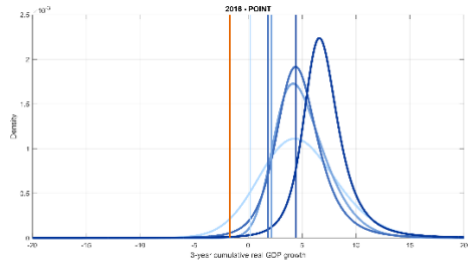
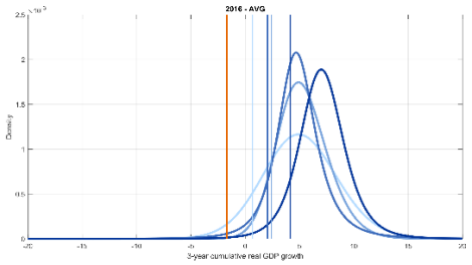
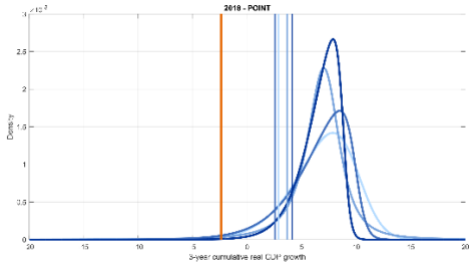
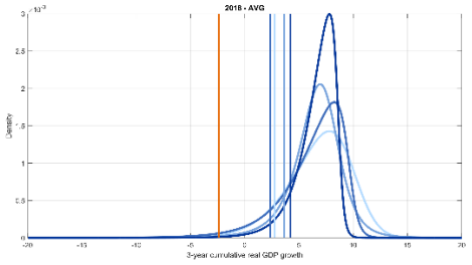
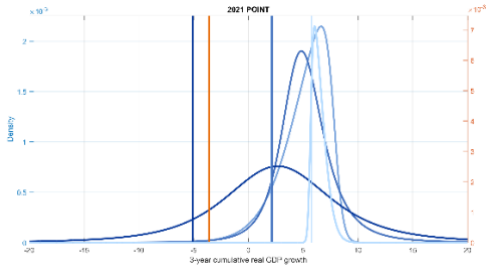
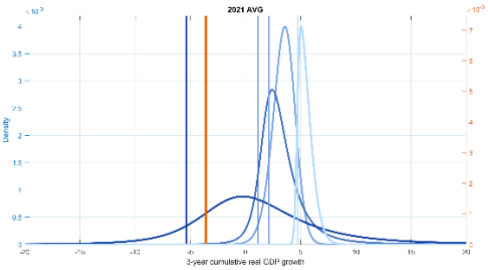
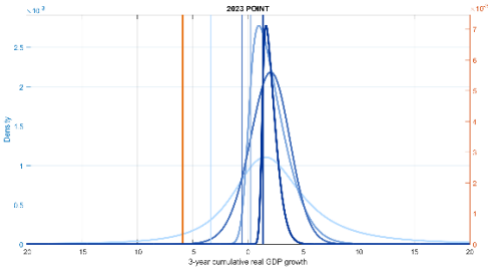
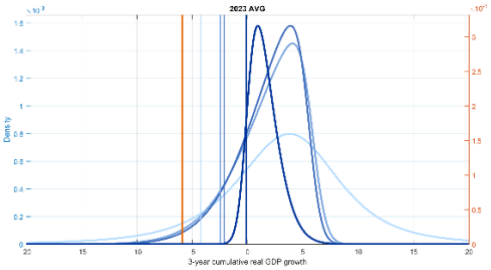
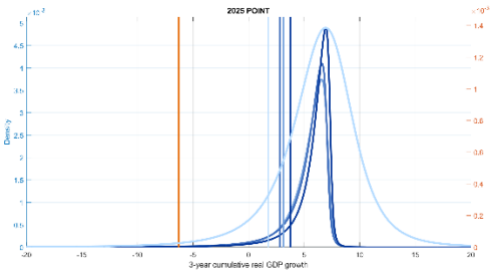
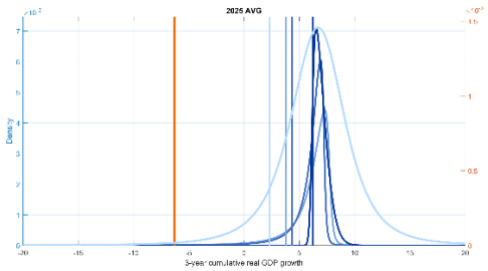
Notes: The table reports summary statistics assessing the in-sample performance of the GaR10 forecasts derived from two model specifications—a basic (small) model including only GDP growth and financial stress (CISS), and a comprehensive (large) model incorporating the full set of macro-financial predictors considered in Table 4. For each forecast horizon (Q1, Q4, and Q12), the hit rate denotes the share of observations where actual year-on-year GDP growth falls below the model-implied 10th percentile. The mean and median errors capture the average and median absolute distance between realized growth and the GaR10 forecast. RMSE refers to the root mean squared error. All metrics are based on fitted values obtained using coefficients estimated over the sample period ending in 2024, and reflect in-sample performance.

Table 6. Evaluation of recession-prediction performance

	AUC				Brier score				Hit rate **			
Horizon	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS
1-quarter ahead	0.87	0.83	0.84	0.81	0.10	0.10	0.25	0.25	0.52	0.52	0.96	0.97
									0.72	0.80	0.84	0.66
4-quarter ahead	0.71	0.68	0.47	0.72	0.11	0.11	0.34	0.28	0.27	0.23	0.77	0.96
									0.64	0.73	1.00	0.96
12-quarter ahead	0.61	0.79	0.43	0.51	0.10	0.09	0.35	0.33	0.11	0.11	0.68	0.74
									0.84	0.89	0.63	0.26
	**False alarm rate ** **				**Median lead (first) ** **				**Coverage window ** **			
	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS
1-quarter ahead	0.05	0.05	0.66	0.65	0.00	0.00	5.00	5.00	0.00	0.00	0.80	1.00
	0.06	0.19	0.24	0.15	1.00	3.50	5.00	3.50	0.20	0.40	0.20	0.40
4-quarter ahead	0.08	0.07	0.69	0.65	3.50	3.50	6.00	5.00	0.40	0.40	0.60	1.00
	0.26	0.38	0.85	0.59	5.00	5.00	5.75	5.00	0.60	0.80	0.80	1.00
12-quarter ahead	0.00	0.06	0.70	0.70	0.00	0.00	6.00	5.00	0.00	0.00	0.40	1.00
	0.53	0.36	0.54	0.07	6.00	6.00	6.00	3.50	0.80	0.60	0.20	0.40
	Always on **				**Max run** **				**Mean share** **			
	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS
1-quarter ahead	0.00	0.00	0.20	0.60	0.00	0.00	2.80	4.20	0.00	0.00	0.53	0.87
	0.00	0.00	0.00	0.20	0.20	0.80	0.20	1.40	0.03	0.20	0.07	0.23
4-quarter ahead	0.00	0.00	0.40	0.60	0.60	0.60	3.00	4.20	0.13	0.13	0.50	0.87
	0.20	0.20	0.40	0.60	2.00	2.80	4.20	4.20	0.33	0.53	0.73	0.87
12-quarter ahead	0.00	0.00	0.40	0.60	0.00	0.00	2.40	4.20	0.00	0.00	0.40	0.90
	0.20	0.40	0.20	0.00	2.60	2.60	1.20	0.60	0.60	0.43	0.20	0.13

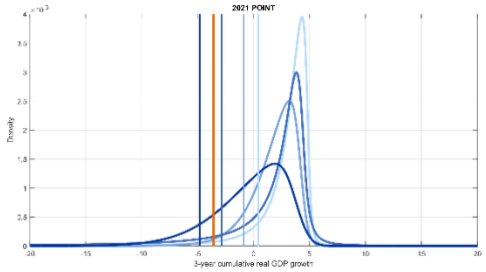
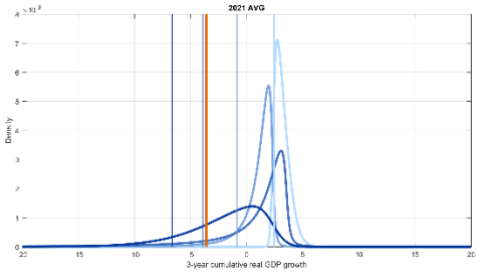
Notes: AUC is the probability that a recession quarter receives a higher predicted probability than a non-recession quarter. The Brier score is the mean squared error between predicted probabilities and realized outcomes (lower is better). The hit rate is the share of recessions correctly signalled, the false alarm rate the share of expansions incorrectly flagged. Median lead (first) measures the median number of quarters between the first correct signal and the onset of a recession, within a six-quarter pre-recession window. Coverage window is the share of that window with at least one correct signal. "Always on" indicates the share of time alerts are active outside recessions; "max run" the longest uninterrupted sequence of alerts; and "mean share" the average fraction of the pre-recession window occupied by alerts. Metrics marked with ** are reported for both the baseline threshold τ=0.30 and the optimal τ^*, where τ^* is calibrated by maximising Youden's J statistic. All metrics are based on fitted values obtained using coefficients estimated over the sample period ending in 2024, and reflect in-sample performance.

Figure 1. Probability distribution of 3-year cumulative real GDP growth

Cutoff: Q4	Cutoff: Q1-Q3
	
	
	
	
	

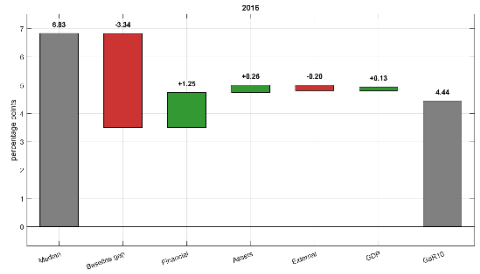
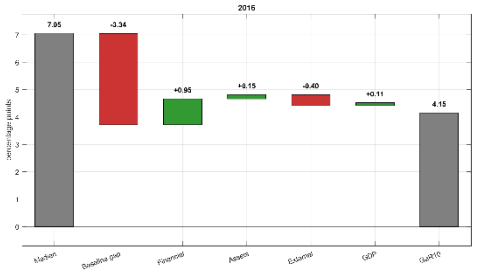
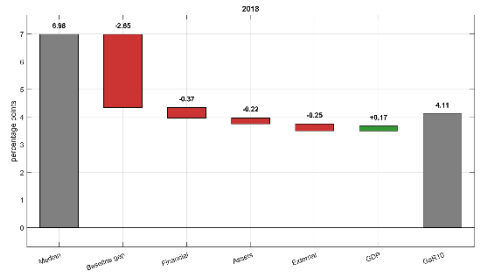
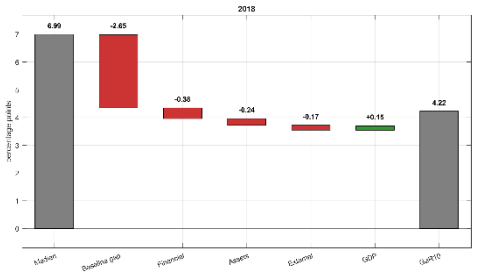
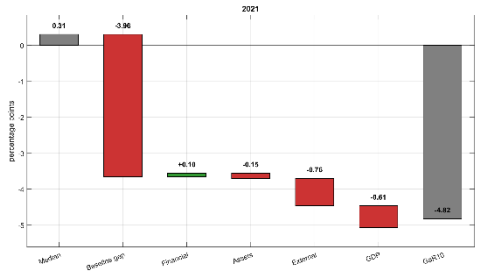
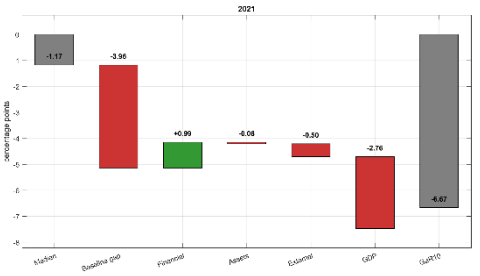
Notes: The forecast distributions are based on an out-of-sample projection, with the cutoff date set at the last quarter (Q4, left column) or the average of the first three quarters (Q1-Q3, right column) of the year preceding the EBA stress test exercise. The figure illustrates the probability distribution of three-year cumulative GDP growth across four different models. These distributions are derived from successive GaR model specifications of increasing complexity—from a parsimonious setup using only GDP and financial turbulence (light blue) to more comprehensive models incorporating the full set of macrofinancial indicators (dark blue). The vertical blue bars denote the GaR10 levels implied by each model, while the vertical orange bar marks the severity of the corresponding EBA adverse scenario.

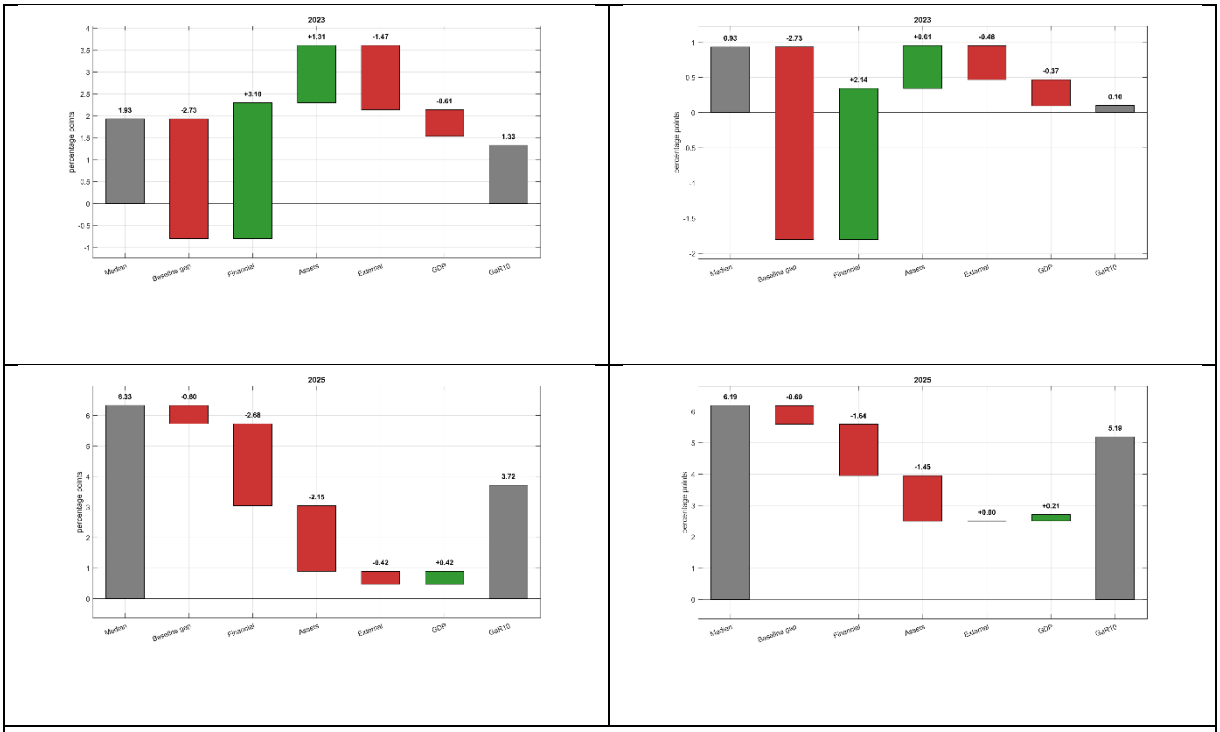
Figure 2. Probability distribution of 3-year cumulative real GDP growth – Nonlinear model specification

Cutoff: Q4	Cutoff: Q1-Q3
	

Notes: The figure reports the conditional probability distributions of three-year cumulative real GDP growth under the nonlinear GaR specification for the 2021 round, with the cutoff date set at Q4 (left panel) and Q1–Q3 (right panel). The distributions incorporate state-dependent and threshold effects in both output and financial stress variables. These distributions are derived from successive GaR model specifications of increasing complexity—from a parsimonious setup using only GDP and financial turbulence (light blue) to more comprehensive models incorporating the full set of macrofinancial indicators (dark blue). The vertical blue bars denote the GaR10 levels implied by each model, while the vertical orange bar marks the severity of the corresponding EBA adverse scenario.

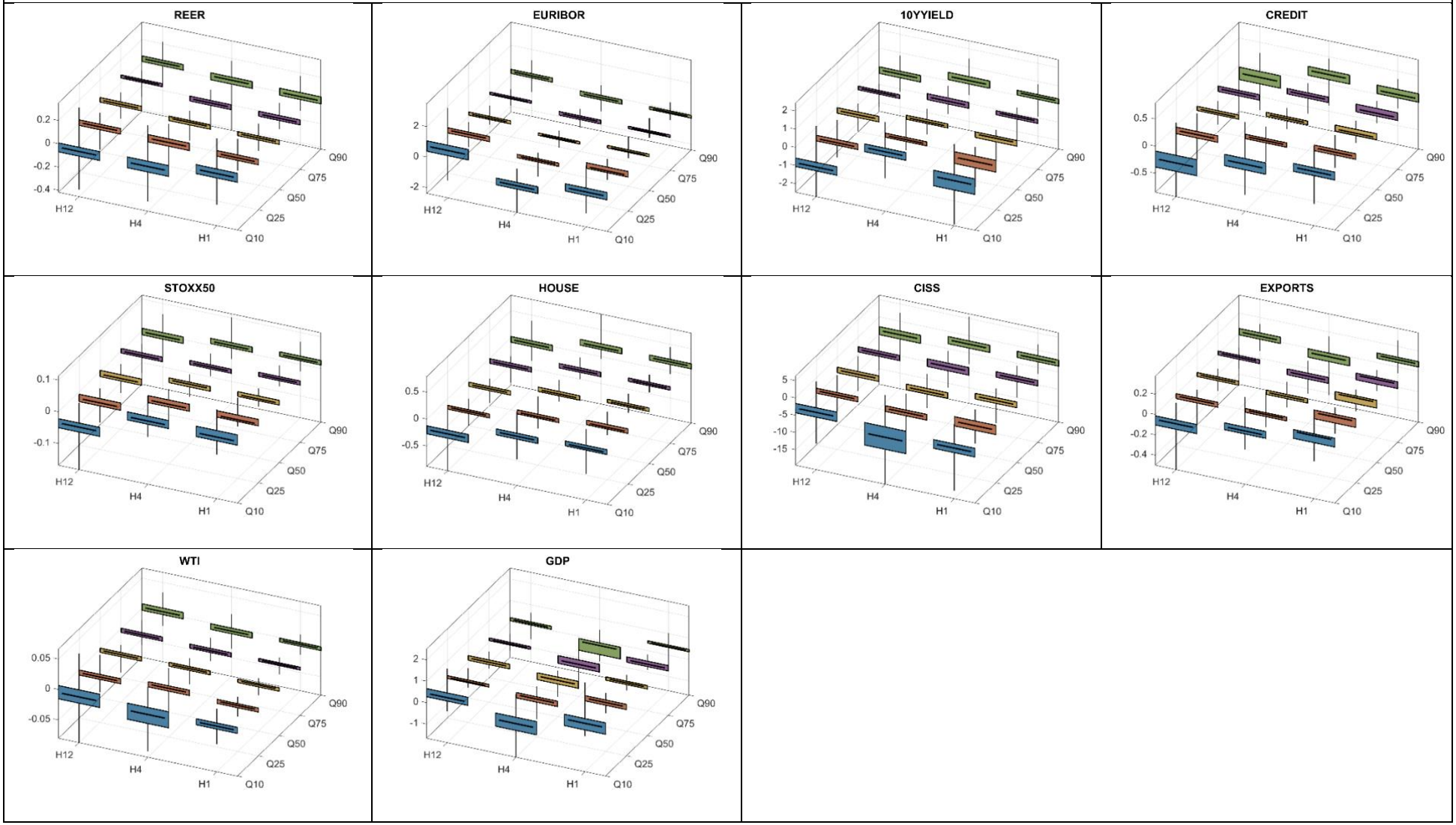
Figure 3. Decomposition of downside growth risks

Cutoff: Q4	Cutoff: Q1-Q3
	
	
	



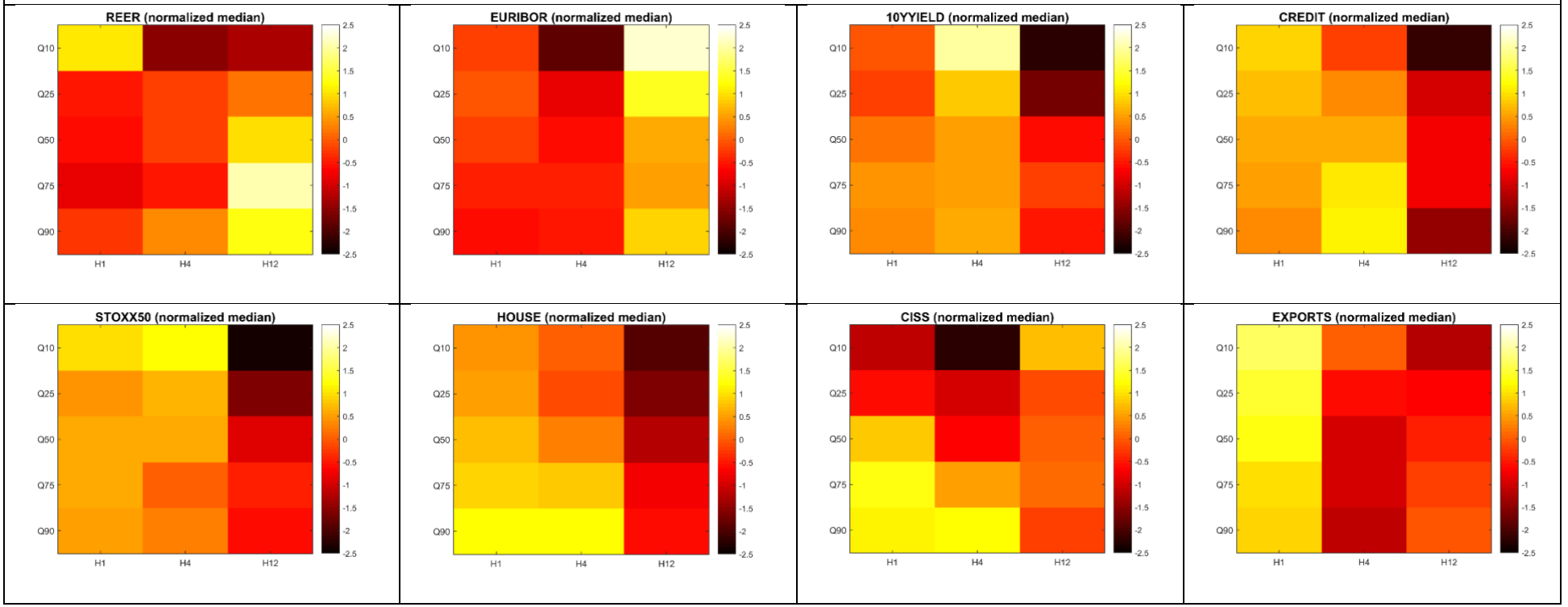
Notes: Counterfactual decomposition of the GaR10–median gap (q10–q50) into contributions from groups of predictors. Bars measure the change in downside risk relative to a baseline re-centred at the median of the last two years. Grey bars mark the initial median and final GaR10 levels; red (green) blocks indicate factors that increase (reduce) downside risks. Factors are grouped as: financial conditions (CISS, short- and long-term rates, credit to the private non-financial sector), asset valuations (EuroStoxx50, house prices), external environment (exports, REER, WTI), and real activity. The difference between start and end bars equals the observed GaR10–median gap at the cutoff time.

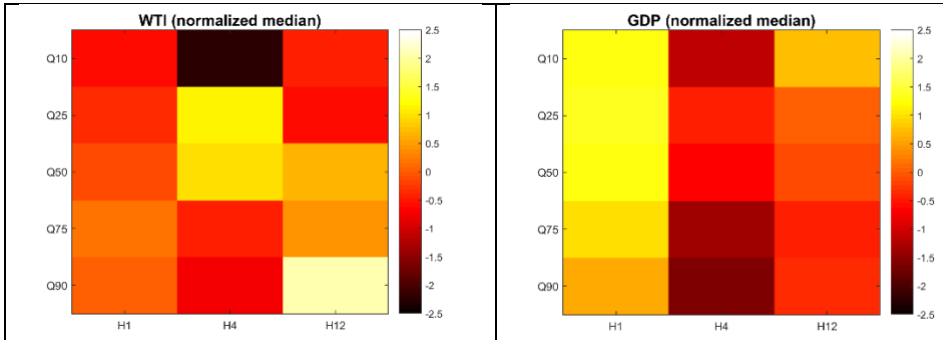
Figure 4. Model and parameter uncertainty: boxplots of coefficient estimates by quantile and forecast horizon



Notes: Estimated coefficients across 1-, 4-, and 12-quarter forecast horizons (x-axis) and conditional quantiles of output growth (y-axis). The boxplots reflect model uncertainty from the alternative regression models, from smallest to largest, as well as parameter uncertainty from 1000 bootstrapped replication of the estimated regression coefficients. The estimation period is Q1:1980-Q4:2024.

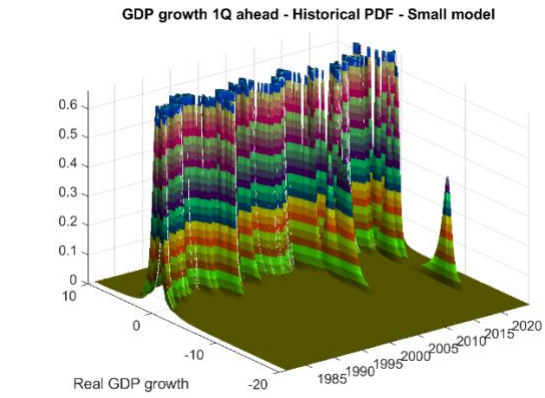
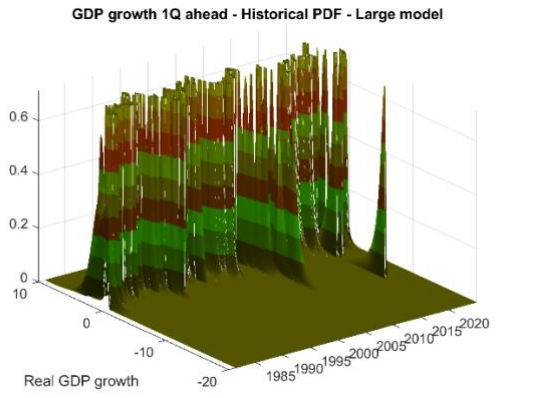
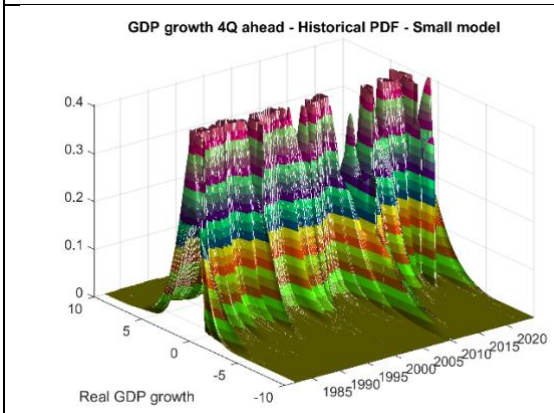
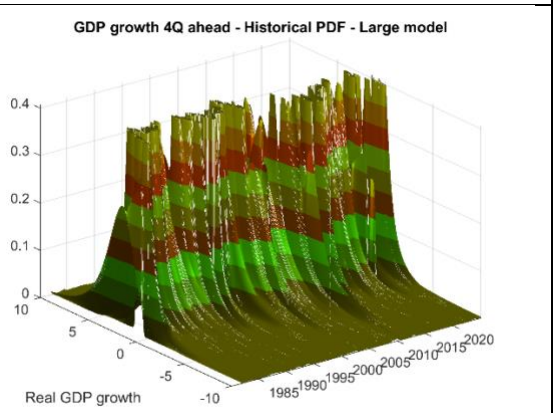
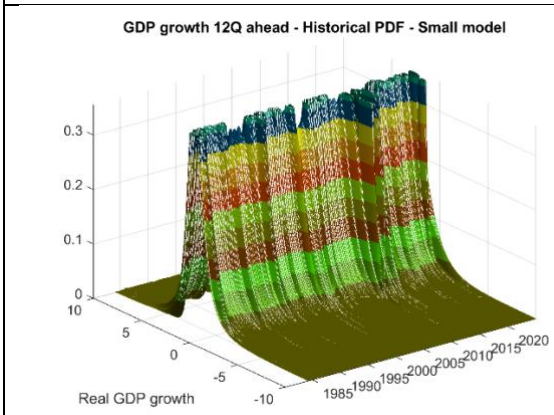
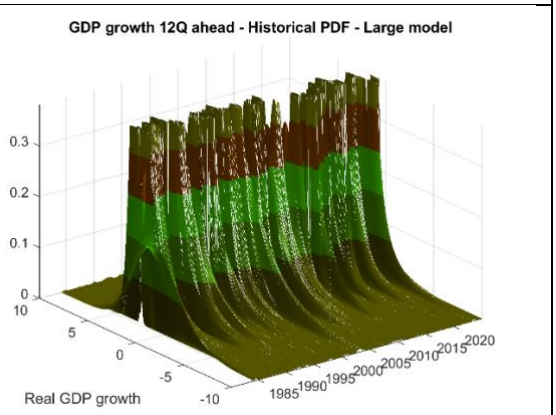
Figure 5. Model and parameter uncertainty: normalized median coefficient estimates by quantile and forecast horizon





Notes: Median coefficient estimate from 1,000 bootstrap replications, normalized by variable to express deviations from the variable-specific mean in standard deviation units (z-scores). Estimates are shown across forecast horizons (x-axis: 1, 4, and 12 quarters) and conditional quantiles of output growth (y-axis). Colors indicate deviations from each variable's **own** average effect; for signs in levels, see Table 4. The estimation sample spans Q1:1980–Q4:2024 and includes multiple model specifications capturing different sets of predictors.

Figure 6. Historical conditional probabilities of GDP growth

Small model	Large model
	
	
	

Notes: Fitted conditional probability density function of future GDP growth one-, four- and twelve-quarter ahead in each year through the estimation sample, using both the small (left) and large (right) GaR models. Sample ends in Q4:2024.

Appendix

Table 1 Appendix. Quantile regressions coefficients across alternative models and forecast horizons - Pre-pandemic sample.												
1-quarter ahead												
	Model #3			**Model #5**			**Model #7**			**Model #10**		
Explanatory variable	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**
REER	-0.02	-0.03	-0.06	0.02	0.00	-0.02	-0.02	0.00	-0.05	-0.01	0.00	-0.01
	0.43	0.28	0.05	0.26	0.95	0.42	0.11	0.92	0.07	0.49	0.79	0.35
Short-term rate	0.35	0.61	0.21	0.05	0.39	-0.24	-0.05	0.11	-0.20	-0.27	-0.25	-0.24
	0.01	0.00	0.33	0.68	0.05	0.18	0.54	0.44	0.20	0.00	0.01	0.01
Long-term rate	-0.45	-0.92	0.29	-0.18	-0.62	0.31	0.04	0.09	0.20	0.28	0.30	0.24
	0.00	0.00	0.19	0.14	0.00	0.09	0.65	0.58	0.23	0.00	0.01	0.01
Private sector credit				0.24	0.23	0.32	0.15	0.18	0.17	0.01	0.05	0.03
				0.00	0.00	0.00	0.00	0.00	0.00	0.71	0.27	0.46
Euro Stoxx 50				0.04	0.04	0.01	0.02	0.01	0.00	0.01	0.01	0.01
				0.00	0.00	0.13	0.00	0.25	0.91	0.00	0.02	0.06
House prices							0.14	0.10	0.08	0.07	0.09	0.02
							0.00	0.04	0.12	0.00	0.00	0.55
CISS							-5.00	-7.41	-2.41	-1.99	-4.48	-0.76
							0.00	0.00	0.03	0.00	0.00	0.26
Exports										0.02	0.04	-0.02
										0.21	0.11	0.45
Oil price (WTI)										0.00	-0.01	0.00
										0.18	0.06	0.92
GDP										0.69	0.54	0.85
										0.00	0.00	0.00

4-quarters ahead												
	Model #3			Model #5			Model #7			Model #10		
Explanatory variable	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**
REER	-0.03	-0.05	-0.02	0.00	-0.03	0.02	-0.05	-0.10	0.02	-0.08	-0.08	-0.06
	0.13	0.19	0.57	0.95	0.41	0.62	0.01	0.01	0.58	0.00	0.10	0.08
Short-term rate	-0.17	-0.58	0.13	-0.35	-0.85	-0.14	-0.41	-0.27	-0.27	-0.44	-0.31	-0.34
	0.23	0.02	0.53	0.01	0.00	0.50	0.00	0.23	0.18	0.00	0.22	0.08
Long-term rate	0.23	0.72	-0.35	0.36	0.96	0.01	0.55	0.35	0.16	0.49	0.40	0.07
	0.13	0.00	0.11	0.02	0.00	0.97	0.00	0.14	0.45	0.00	0.14	0.75
Private sector credit				0.10	0.12	0.26	0.03	-0.04	0.22	0.02	-0.07	0.11
				0.02	0.09	0.00	0.51	0.61	0.00	0.76	0.49	0.18
Euro Stoxx 50				0.03	0.03	0.00	0.00	0.01	0.00	0.01	0.02	0.01
				0.00	0.00	0.94	0.53	0.49	0.83	0.17	0.26	0.32
House prices							0.10	0.04	0.06	0.12	0.08	0.11
							0.01	0.53	0.38	0.00	0.30	0.09
CISS							-5.29	-11.41	-0.79	-4.63	-10.30	-1.77
							0.00	0.00	0.58	0.00	0.00	0.25
Exports										-0.06	0.03	-0.10
										0.06	0.65	0.04
Oil price (WTI)										-0.01	-0.02	0.00
										0.16	0.04	0.95
GDP										0.18	-0.12	0.24
										0.22	0.68	0.30

12-quarters ahead												
	Model #3			Model #5			Model #7			Model #10		
Explanatory variable	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**	**OLS**	**10th pct**	**90th pct**
REER	0.04	-0.01	0.04	0.06	0.08	0.04	0.03	0.01	0.04	0.01	-0.01	0.04
	0.08	0.70	0.19	0.01	0.05	0.02	0.23	0.76	0.07	0.83	0.84	0.10
Short-term rate	0.08	0.06	0.31	0.21	0.67	-0.03	0.26	0.18	0.14	0.19	0.27	0.19
	0.57	0.80	0.14	0.14	0.01	0.89	0.07	0.39	0.51	0.19	0.27	0.40
Long-term rate	-0.24	-0.35	-0.07	-0.44	-0.85	0.06	-0.42	-0.59	-0.08	-0.42	-0.85	-0.20
	0.11	0.12	0.74	0.00	0.00	0.79	0.01	0.01	0.73	0.01	0.00	0.39
Private sector credit				-0.22	-0.33	-0.02	-0.17	-0.11	0.01	-0.28	-0.39	-0.10
				0.00	0.00	0.83	0.00	0.19	0.95	0.00	0.00	0.30
Euro Stoxx 50				0.01	-0.03	0.03	-0.01	-0.04	0.02	-0.01	-0.04	0.02
				0.32	0.03	0.01	0.36	0.00	0.15	0.37	0.00	0.14
House prices							-0.08	-0.11	-0.04	-0.12	-0.22	-0.03
							0.08	0.11	0.55	0.01	0.01	0.69
CISS							-3.02	-4.78	-2.51	-1.05	-1.93	-1.23
							0.00	0.00	0.11	0.35	0.31	0.48
Exports										-0.08	-0.17	-0.01
										0.02	0.00	0.85
Oil price (WTI)										0.00	0.01	0.00
										0.65	0.59	0.95
GDP										0.59	0.91	0.24
										0.00	0.00	0.35

Notes: Note: Coefficients and p-values are based on OLS and quantile regressions of growth distributions at the 10th and 90the percentiles. The estimation period is Q1:1980-Q4:2019.

Table 2 Appendix. Model fit statistics for GaR10 forecasts at different horizons

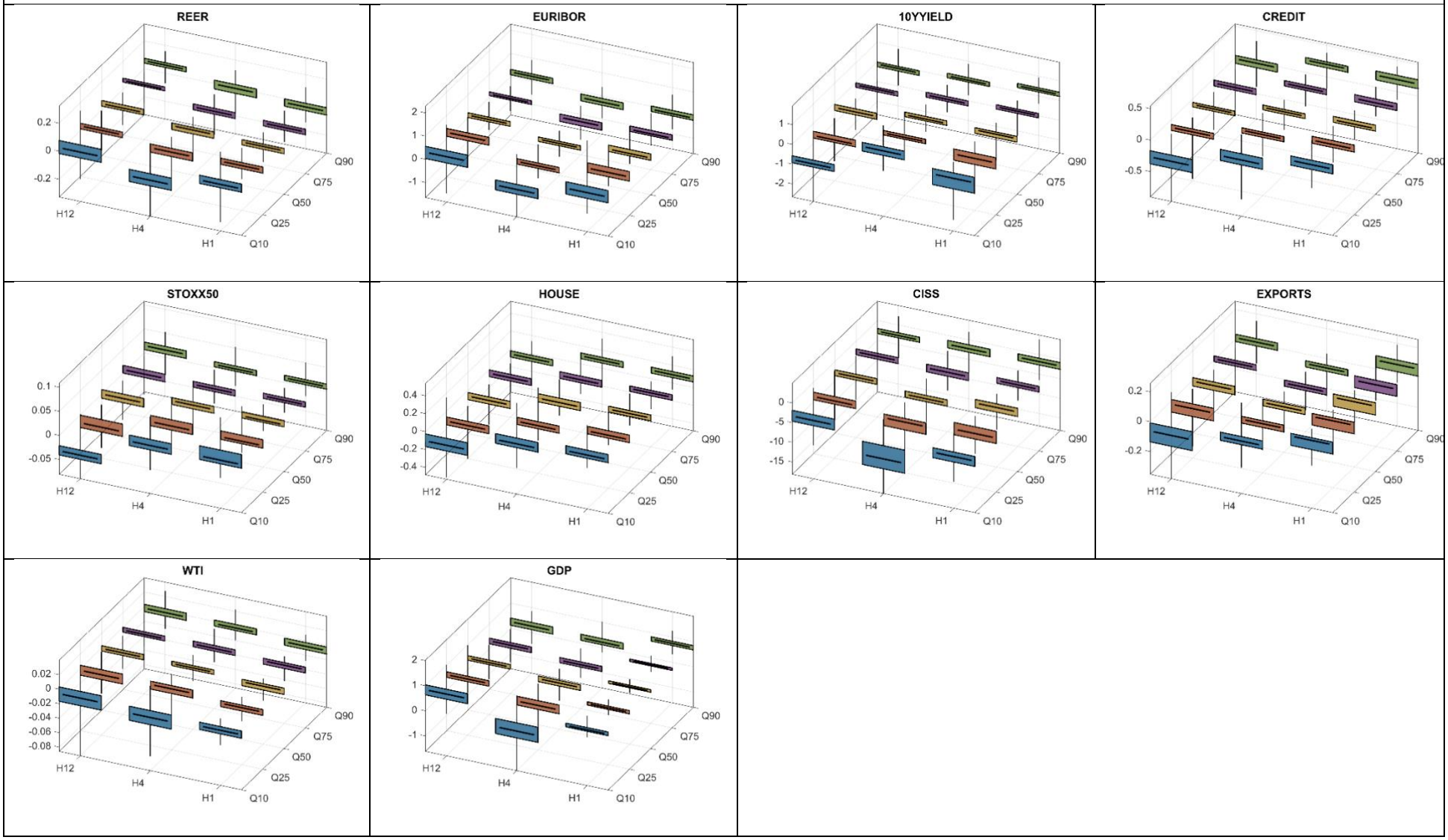
Horizon	Hit rate		Mean error		Median error		RMSE	
	Small	Large	Small	Large	Small	Large	Small	Large
1-quarter ahead	9.1%	7.4%	0.74	0.69	0.69	0.63	1.06	0.95
4-quarter ahead	8.7%	7.0%	1.31	1.62	1.08	1.38	2.06	2.29
12-quarter ahead	8.5%	6.7%	1.58	1.42	1.92	1.02	2.29	2.22

Notes: The table reports summary statistics assessing the in-sample performance of the GaR10 forecasts derived from two model specifications—a basic (small) model including only GDP growth and financial stress (CISS), and a comprehensive (large) model incorporating the full set of macro-financial predictors considered in Table 4. For each forecast horizon (Q1, Q4, and Q12), the hit rate denotes the share of observations where actual year-on-year GDP growth falls below the model-implied 10th percentile. The mean and median errors capture the average and median absolute distance between realized growth and the GaR10 forecast. RMSE refers to the root mean squared error All metrics are based on fitted values obtained using coefficients estimated over the sample period ending in Q4:2019, and reflect in-sample performance

Table 3 Appendix. Evaluation of recession-prediction performance												
	AUC				**Brier score**				**Hit rate ****			
Horizon	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS
1-quarter ahead	0.91	0.91	0.86	0.83	0.09	0.09	0.24	0.24	0.52	0.52	0.96	0.96
									0.83	0.87	0.83	0.67
4-quarter ahead	0.77	0.70	0.46	0.75	0.11	0.11	0.34	0.27	0.30	0.25	0.70	0.96
									0.70	0.60	1.00	0.58
12-quarter ahead	0.43	0.91	0.43	0.50	0.10	0.08	0.35	0.33	0.12	0.29	0.65	0.71
									0.29	1.00	0.18	0.29
	False alarm rate **				**Median lead (first) ****				**Coverage window ****			
	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS
1-quarter ahead	0.04	0.04	0.66	0.65	0.00	0.00	4.00	5.50	0.00	0.00	0.75	1.00
	0.13	0.12	0.20	0.11	3.50	3.50	0.00	3.50	0.50	0.50	0.00	0.50
4-quarter ahead	0.08	0.05	0.70	0.65	3.50	3.00	4.00	5.50	0.50	0.50	0.50	1.00
	0.25	0.24	0.88	0.17	5.00	4.67	5.67	3.50	0.75	0.75	0.75	0.50
12-quarter ahead	0.00	0.06	0.71	0.70	0.00	0.00	6.00	5.50	0.00	0.00	0.50	1.00
	0.10	0.21	0.09	0.08	3.50	6.00	6.00	3.50	0.50	0.75	0.25	0.50
	Always on**				**Max run****				**Mean share****			
	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS	Small GaR	Large GaR	Univariate GDP	Univariate CISS
1-quarter ahead	0.00	0.00	0.00	0.75	0.00	0.00	2.00	4.00	0.00	0.00	0.42	0.92
	0.00	0.00	0.00	0.00	0.75	0.75	0.00	1.00	0.17	0.17	0.00	0.25
4-quarter ahead	0.00	0.00	0.25	0.75	0.75	0.50	2.00	4.00	0.17	0.13	0.33	0.92
	0.25	0.25	0.25	0.25	2.50	2.50	3.75	1.75	0.42	0.42	0.67	0.29
12-quarter ahead	0.00	0.00	0.25	0.75	0.00	0.00	1.75	4.00	0.00	0.00	0.29	0.92
	0.00	0.00	0.25	0.00	1.00	1.75	1.50	0.75	0.21	0.29	0.25	0.17

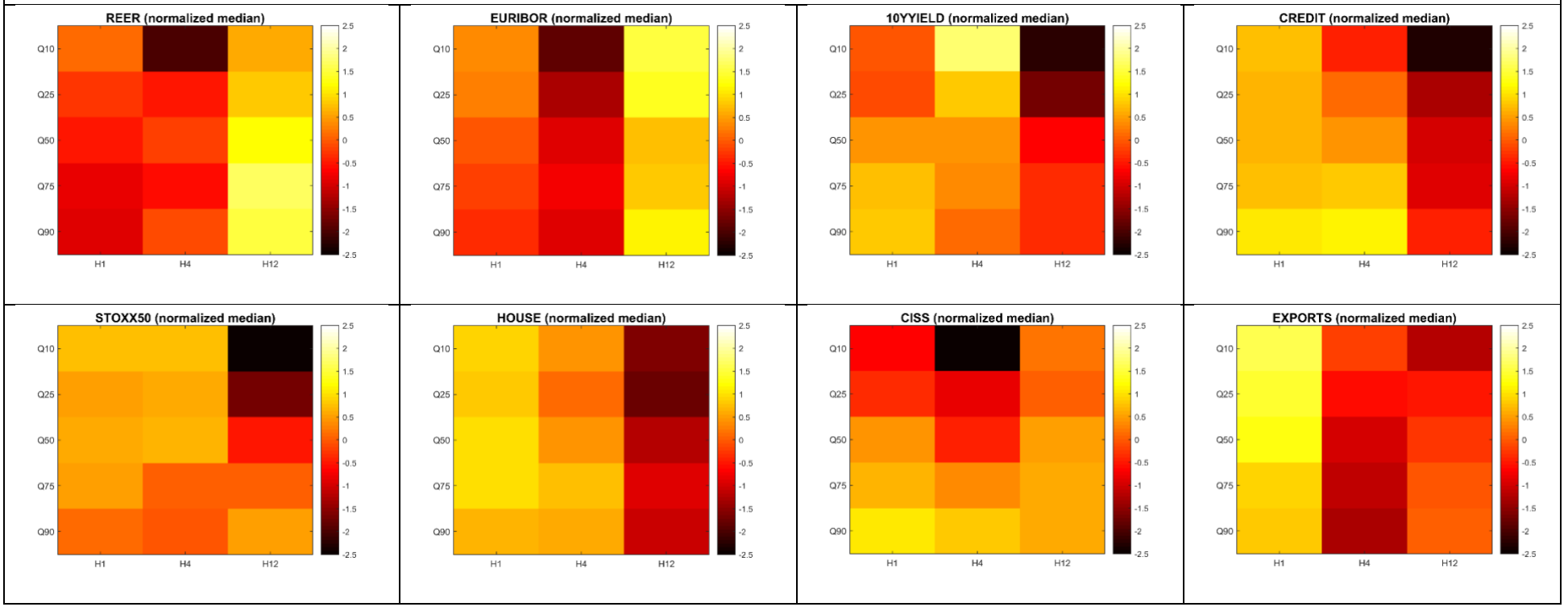
Notes: AUC is the probability that a recession quarter receives a higher predicted probability than a non-recession quarter. The Brier score is the mean squared error between predicted probabilities and realized outcomes (lower is better). The hit rate is the share of recessions correctly signalled, the false alarm rate the share of expansions incorrectly flagged. Median lead (first) measures the median number of quarters between the first correct signal and the onset of a recession, within a six-quarter pre-recession window. Coverage window is the share of that window with at least one correct signal. "Always on" indicates the share of time alerts are active outside recessions; "max run" the longest uninterrupted sequence of alerts; and "mean share" the average fraction of the pre-recession window occupied by alerts. Metrics marked with ** are reported for both the baseline threshold τ=0.30 and the optimal τ^*, where τ^* is calibrated by maximising Youden's J statistic. All metrics are based on fitted values obtained using coefficients estimated over the sample period ending in Q4:2019, and reflect in-sample performance

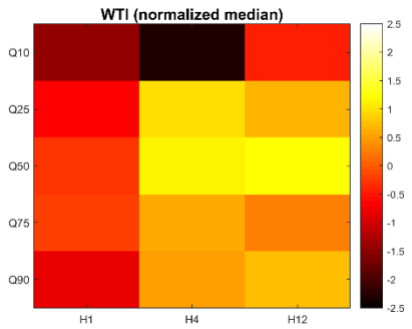
Figure 1 Appendix. Model and parameter uncertainty: boxplots of coefficient estimates by quantile and forecast horizon

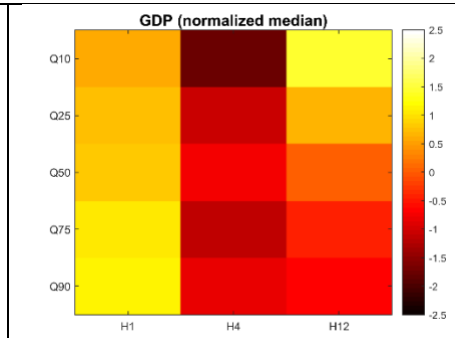


Notes: Estimated coefficients across 1-, 4-, and 12-quarter forecast horizons (x-axis) and conditional quantiles of output growth (y-axis). The boxplots reflect model uncertainty from the alternative regression models, from smallest to largest, as well as parameter uncertainty from 1000 bootstrapped replication of the estimated regression coefficients. The estimation period is Q1:1980-Q4:2019.

Figure 2 Appendix. Model and parameter uncertainty: normalized median coefficient estimates by quantile and forecast horizon

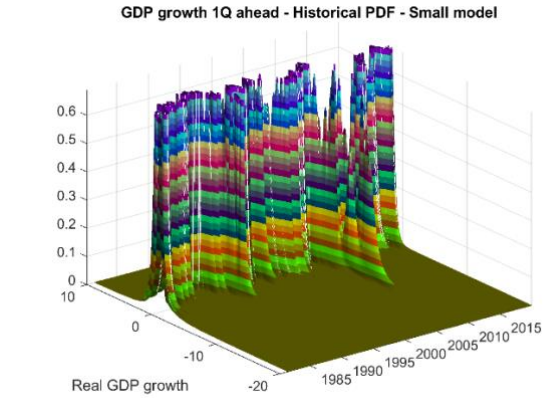
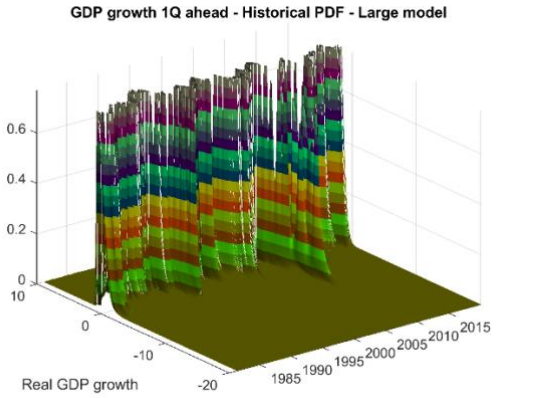
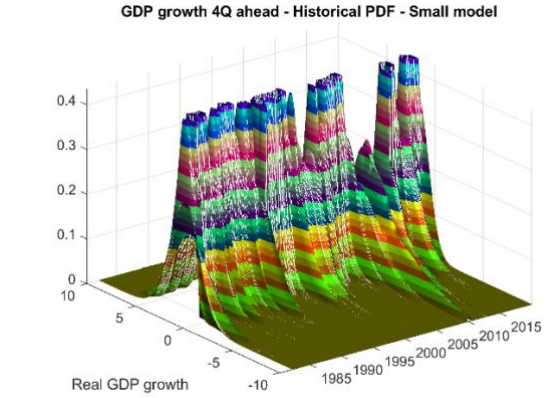
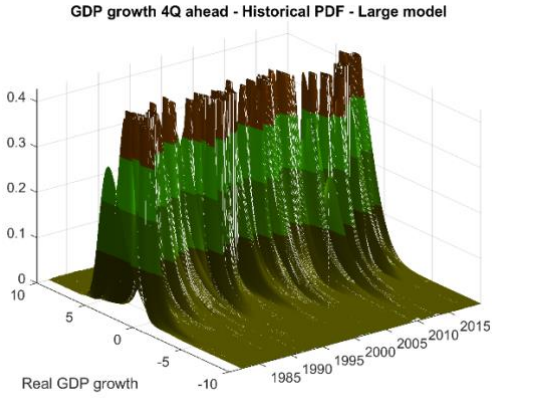
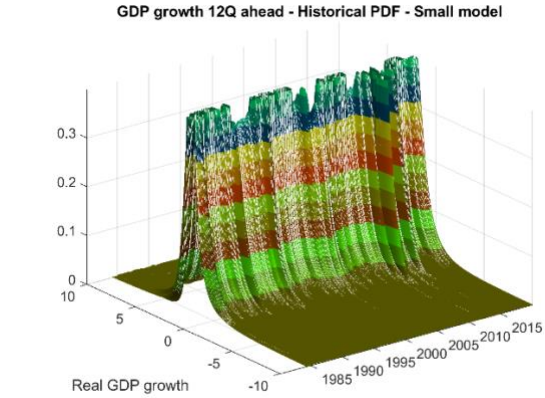
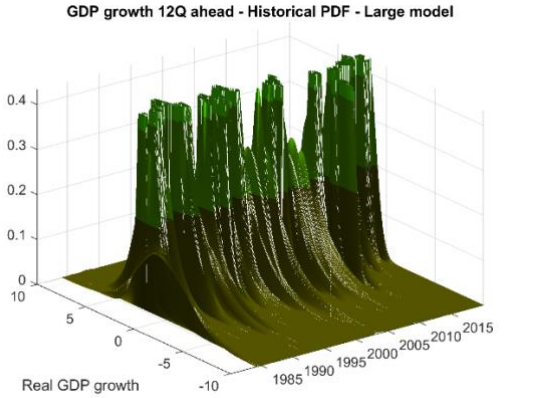






Notes: Median coefficient estimate from 1,000 bootstrap replications, normalized by variable to express deviations from the variable-specific mean in standard deviation units (z-scores). Estimates are shown across forecast horizons (x-axis: 1, 4, and 12 quarters) and conditional quantiles of output growth (y-axis). Colors indicate deviations from each variable's **own** average effect; for signs in levels, see Table 1 Appendix. The estimation sample spans Q1:1980–Q4:2019 and includes multiple model specifications capturing different sets of predictors.

Figure 3 Appendix. Historical conditional probabilities of GDP growth

Small model	Large model
	
	
	

Notes: Fitted conditional probability density function of future GDP growth one-, four- and twelve-quarter ahead in each year through the estimation sample, using both the small (left) and large (right) GaR models. Sample ends in Q4:2019.

HOW SEVERE ARE EUROPEAN REGULATORY STRESS TEST SCENARIOS?

A PROBABILISTIC CALIBRATION FOR THE EURO AREA

European Investment Bank

pdf: ISBN 978-92-861-6088-2

2025-0384