Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Does Outcome Measurement Matter for Forecasting? Evidence from Fiscal Vulnerability in Developing Economies

Sonam Tobgye

sonam.tobgye2@rmit.edu.au

RMIT University

Research Article

Keywords: Fiscal vulnerability, Macroeconomic forecasting, Persistence, Machine learning, Fiscal surveillance, Developing economies

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11476920/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Does Outcome Measurement Matter for Forecasting? Evidence from Fiscal Vulnerability in Developing Economies

Abstract

The construction of multidimensional outcome measures can influence the evaluation and interpretation of forecasting models, yet this sensitivity is often overlooked in empirical applications. This study examines how alternative definitions of fiscal vulnerability affect predictive performance, incremental forecasting value beyond persistence, and surveillance conclusions in developing economies. Using an annual panel of 127 developing economies over 1982–2024, three vulnerability measures are constructed from eight common fiscal, macroeconomic, and external indicators, capturing the breadth of adverse conditions, the severity of threshold exceedances, and a principal component-based measure. Forecasting models are assessed using chronological validation and out-of-sample evaluation against logistic regression and fitted persistence benchmarks. In the 2018–2023 external evaluation, the selected model achieves AUCs of 0.819, 0.813, and 0.959 for breadth, severity, and PCA, respectively. However, incremental AUC gains over persistence are statistically significant only for breadth (+0.022) and severity (+0.020), with no corresponding improvement established for PCA. Permutation importance identifies fiscal balance and government debt as leading predictors, while the contributions of other indicators differ across definitions. Breadth-based forecasts also provide additional information about new vulnerability onsets, and country-level surveillance rankings vary across measures. These findings demonstrate that outcome measurement affects conclusions about forecasting performance, predictor relevance, and surveillance priorities, while high apparent predictability does not necessarily translate into incremental predictive value.

Keywords: Fiscal vulnerability; Macroeconomic forecasting; Persistence; Machine learning; Fiscal surveillance; Developing economies

JEL classification: C45; C53; E62; H63; O19

1. Introduction

Fiscal vulnerability remains an important concern for developing economies, where limited revenue capacity, macroeconomic instability, and external financing constraints can restrict governments' ability to respond to adverse shocks. The sustainability of public finances depends not only on public debt and fiscal balances but also on economic growth, revenue mobilisation, debt composition, and external financing conditions (Kraay and Nehru, 2006; Bi et al., 2016; Jarmuzek and Vesperoni, 2018). Fiscal distress may therefore arise from the interaction of several adverse conditions rather than the deterioration of a single indicator. Evidence from fiscal-distress early-warning models further suggests that fiscal and non-fiscal variables provide complementary predictive information, supporting a broader approach to assessing macro-fiscal vulnerabilities (Bruns and Poghosyan, 2018).

The measurement of fiscal vulnerability is particularly relevant for fiscal surveillance because vulnerability is not synonymous with realised fiscal or financial crises. Sovereign-debt, banking, and currency crises represent episodes of acute distress, whereas underlying macroeconomic and fiscal weaknesses may accumulate or persist without culminating in a formally recorded crisis. Macroeconomic fundamentals, external liabilities, and financing conditions play important roles in sovereign and external crises, while banking-sector instability can interact with sovereign risk and amplify financial pressures (Kaminsky and Reinhart, 1999; Catão and Milesi-Ferretti, 2014; Fiordelisi et al., 2020). Multidimensional vulnerability measures therefore provide a complementary surveillance perspective by identifying adverse fiscal, macroeconomic, and external conditions beyond realised crisis episodes. Indicator-based approaches have similarly been used to distinguish different degrees of fiscal vulnerability rather than relying exclusively on observed crisis events (Stoian et al., 2018).

A substantial literature has examined the forecasting of sovereign-debt and financial crises using macroeconomic and financial indicators. Early-warning models have traditionally employed statistical approaches to identify conditions associated with subsequent crisis episodes (Kaminsky et al., 1998; Bussière and Fratzscher, 2006; Dawood et al., 2017). More recently, machine-learning methods have been applied to accommodate nonlinear relationships and interactions among predictors, with evidence of improved forecasting performance in some financial-crisis applications (Fioramanti, 2008; Liu et al., 2022; Bluwstein et al., 2023; Du Plessis and Fritsche,

2025). Nevertheless, greater model flexibility does not consistently translate into superior out-of-sample performance, and conventional logistic regression can remain competitive under recursive evaluation (Beutel et al., 2019). These findings highlight the importance of evaluating forecasting gains against appropriate benchmarks rather than interpreting high predictive accuracy as evidence of model superiority.

A further issue arises when the forecasting target is itself a constructed measure. Unlike realised crisis events, multidimensional fiscal vulnerability depends on the selection, scaling, and aggregation of underlying indicators. Fiscal-vulnerability assessments have employed alternative approaches to characterise fiscal weakness and exposure to adverse economic conditions (Stoian et al., 2018; Rajaguru et al., 2021), while research on multidimensional vulnerability measurement highlights the influence of indicator selection and aggregation on the interpretation of composite measures (Adger, 2006; Anderson et al., 2019). Alternative constructions may consequently capture different aspects of the same economic conditions and produce different vulnerability classifications. Breadth-based measures emphasise the simultaneous occurrence of adverse conditions, severity measures incorporate the magnitude of threshold exceedances, and principal-component approaches summarise dominant statistical variation across indicators. Related factor-based methods have been applied to financial-crisis forecasting and the construction of financial-vulnerability indicators (Cipollini and Kapetanios, 2009; Kuek et al., 2020). These measurement differences may also alter the incidence and persistence of the outcome being forecast, even when the underlying economic indicators remain unchanged.

Persistence introduces a related challenge. Fiscal positions, debt burdens, and macroeconomic conditions may exhibit substantial temporal dependence, implying that previous vulnerability can contain information about its subsequent state. Strong out-of-sample performance may therefore reflect the continuation of existing vulnerability rather than incremental information from additional macro-fiscal predictors. This possibility motivates comparisons against persistence-based benchmarks, consistent with established principles of forecast evaluation (De Gooijer and Hyndman, 2006; Hewamalage et al., 2023). In this setting, the relevant question is whether multivariate forecasts improve upon predictions based on the preceding continuous vulnerability score rather than merely outperforming an unconditional forecasting benchmark.

These considerations have implications for early-warning performance and country-level surveillance. A model may achieve strong discrimination by predicting the continuation of existing vulnerability while providing relatively limited information about new vulnerability onsets. The distinction between crisis onset and continuing or post-crisis conditions is well recognised in early-warning research (Bussière and Fratzscher, 2006; Dawood et al., 2017). Moreover, alternative vulnerability definitions may generate similar aggregate forecasting performance while assigning different predicted probabilities and relative rankings to individual economies. A comprehensive assessment should therefore consider whether forecasting gains remain robust across alternative outcome definitions, whether models provide information about new vulnerability episodes, and whether country-level surveillance rankings are stable.

Credible forecasting evaluation further requires preserving the information available when predictions are formed. This is especially important when vulnerability thresholds, scaling parameters, or statistical components are estimated from historical observations, since using subsequent observations during model development can introduce look-ahead bias. Chronological validation, training-sample-based preprocessing, and appropriate benchmark comparisons are important for reliable out-of-sample assessment (Hewamalage et al., 2023). Forecast performance should also distinguish discrimination from probability accuracy, with proper scoring rules providing a principled basis for evaluating probabilistic forecasts (Gneiting and Raftery, 2007).

Against this background, this study addresses two related questions: how sensitive are forecasting performance and surveillance conclusions to alternative constructions of fiscal vulnerability, and to what extent do multivariate models provide predictive information beyond the persistence of the constructed outcome? The analysis uses an annual panel of developing economies constructed from the June 2026 release of the Global Macro Database (Müller et al., 2025). Three vulnerability measures are derived from the same eight fiscal, macroeconomic, and external-sector dimensions: the breadth of simultaneous adverse conditions, the severity of threshold exceedances, and a PCA-based statistical component. Forecasting models are evaluated using expanding-window rolling-origin validation, with vulnerability-measurement parameters and predictor preprocessing estimated from the corresponding historical training observations. The selected multivariate specification is compared with conventional logistic regression and persistence-based benchmarks using separate calibration and out-of-sample evaluation periods. Additional analyses examine

vulnerability onsets, alternative measurement specifications, and country-level surveillance rankings.

The empirical findings reveal a distinction between absolute forecastability and incremental predictive information. Fiscal vulnerability is strongly forecastable across all three definitions, but the contribution of the broader macro-fiscal information set varies considerably. Under the breadth and severity measures, the selected multivariate model improves out-of-sample discrimination and probability accuracy relative to fitted lagged-score persistence. Under the PCA-based definition, however, the previous continuous vulnerability score already provides highly informative forecasts, and the multivariate model offers no statistically distinguishable improvement. The breadth-based specification also retains incremental predictive information for new vulnerability onsets, whereas corresponding gains are weaker under severity and absent under PCA. Country-level surveillance rankings are more closely aligned between breadth and severity than with the PCA-based measure, indicating that alternative constructions of the same underlying vulnerability concept can generate different decision-relevant signals even when evaluated using a common forecasting framework.

This study contributes to the literature on multidimensional measurement and empirical forecast evaluation in three respects. First, it examines the sensitivity of forecasting conclusions to alternative constructions of a multidimensional outcome, holding the underlying indicator set and forecasting framework constant. By comparing breadth-, severity-, and principal-component-based definitions of fiscal vulnerability, the analysis assesses whether differences in outcome construction translate into differences in forecastability and predictive evaluation. Second, it distinguishes absolute predictive performance from incremental information beyond outcome persistence. This distinction is particularly relevant for composite measures whose temporal dependence may vary with their construction, potentially producing high forecasting accuracy without corresponding gains from additional predictors. Third, it extends measurement-sensitivity analysis beyond aggregate performance metrics to the identification of new vulnerability episodes and the relative ranking of economies for surveillance. Taken together, these contributions demonstrate how the operational definition of a multidimensional outcome can influence conclusions about model performance, the incremental value of additional predictors, and decision-oriented forecasting. The study does not introduce a new forecasting algorithm; rather, it

provides an empirical assessment of the robustness of forecasting conclusions to alternative outcome definitions within a common evaluation framework.

The remainder of the paper is organised as follows. Section 2 describes the data, vulnerability measures, and empirical forecasting strategy. Section 3 presents the empirical results, while Section 4 concludes with policy implications and limitations.

2. Data and Empirical Strategy

2.1 Data source and sample definition

The analysis draws on the June 2026 release of the Global Macro Database (Müller et al., 2025), which provides harmonised annual fiscal, monetary, real-sector, and external-sector series for a broad cross-section of economies. The study focuses on countries classified by the GMD as low-income, lower-middle-income, or upper-middle-income, consistent with the paper's interest in fiscal vulnerability in developing economies. High-income economies are excluded because fiscal financing conditions, domestic capital-market depth, monetary credibility, revenue capacity, and exposure to foreign-currency constraints differ systematically from those prevailing in developing economies. Income-group assignments follow the classifications reported in the June 2026 GMD release and are held fixed over the sample period; they therefore serve as sample classifications rather than historical measures of each country's income status. The income-group restriction yields 129 economies and 5,805 country-year observations over 1980-2024. Construction of the vulnerability measures requires contemporaneous information on eight fiscal, macroeconomic, and external dimensions. After these data requirements and the temporal structure of the forecasting exercise are imposed, the prepared analytical panel contains 4,527 country-year observations for 127 economies over 1982-2024. The panel is unbalanced, reflecting differences in the historical availability of individual series across countries. The forecasting design subsequently separates the sample chronologically into a development period, a calibration period, held-out evaluation samples, and a final 2024 sensitivity period, as described in Section 2.2.

Only information that would have been observable at the relevant point in time is retained. Where the GMD identifies observations as forecasts, those values are excluded before constructing growth rates, exchange-rate changes, lagged predictors, vulnerability thresholds, and outcome measures. The sample is capped at 2024 because forecast status is not consistently identified for observations extending beyond that year. The principal external evaluation therefore covers 2018–

2023, while 2024 is retained separately as a sensitivity exercise. Country-year observations are indexed by ISO3 country code and calendar year, and changes in variables are calculated only across consecutive annual observations so that gaps in country coverage do not generate artificial growth or depreciation rates. No interpolation, forward filling, or backward filling is used to complete the historical panel. The baseline fiscal measures comprise government debt, net lending or borrowing, government revenue, and government expenditure as shares of GDP. These are combined with real GDP growth, inflation, the current-account balance, and exchange-rate depreciation to form the eight dimensions underlying the fiscal-vulnerability measures developed in the next subsection. Alternative government-sector definitions are examined separately as robustness checks rather than incorporated into the baseline outcome construction. Detailed sample-construction procedures, variable definitions, source mappings, and data-coverage diagnostics are reported in Supplementary Appendix A (Online Resource 1), Sections A.1-A.2, including Tables A1.1-A1.2 and A2.1-A2.2.

2.2 Empirical strategy

The empirical design is a recursive pseudo-out-of-sample forecasting exercise in which fiscal vulnerability is a generated binary outcome. Let $i = 1, \dots, N$ index countries, $t = 1, \dots, T$ years, and $k = 1, \dots, 8$ the fiscal, macroeconomic, and external dimensions entering the vulnerability measure. Denote the contemporaneous outcome vector by $\mathbf{Z}_{it} = (Z_{1it}, \dots, Z_{8it})'$, and let $d_k \in \{-1, +1\}$ orient each component such that $d_k Z_{kit}$ is increasing in adverse conditions. For the complete outcome observations in the 1982-2005 development sample, $\mathcal{D}_0$, define the adverse cut-off

$$c_k = \begin{cases} Q_{0.75}(Z_k \mid \mathcal{D}_0), & d_k = +1, \\ Q_{0.25}(Z_k \mid \mathcal{D}_0), & d_k = -1. \end{cases}$$

The baseline breadth measure is then

$$R_{kit} = \mathbf{1}\{d_k(Z_{kit} - c_k) \geq 0\}, \qquad B_{it} = \sum_{k=1}^{8} R_{kit}\,, \qquad V_{it}^{B} = \mathbf{1}\{B_{it} \geq 3\}. \quad (1)$$

Thus, V_{it}^{B} identifies country-years in which at least three dimensions simultaneously occupy their adverse historical region. The thresholds are distributional parameters and are not interpreted as structural fiscal limits.

Measurement robustness is evaluated by holding $\mathbf{Z}_{it}$ fixed and changing only the mapping from the underlying macro-fiscal state into the binary outcome. The severity construction replaces the count in (1) with the mean standardized distance beyond the corresponding adverse thresholds:

$$S_{it} = \frac{1}{8}\sum_{k=1}^{8}\left[\frac{d_k(Z_{kit} - c_k)}{\sigma_{k,0}}\right]_{+}, \qquad V_{it}^{S} = \mathbf{1}\{S_{it} \geq s_{0.75}\}, \qquad (2)$$

where $[x]_{+} = \max\{x, 0\}$, $\sigma_{k,0}$ is the development-sample standard deviation of component k, and $s_{0.75} = Q_{0.75}(S \mid \mathcal{D}_0)$. A third outcome is obtained from the first principal component of the same adversity-oriented variables. Let

$$\tilde{Z}_{kit} = \frac{d_k Z_{kit} - \mu_{k,0}}{\sigma_{k,0}^{a}}, \qquad P_{it} = \hat{\boldsymbol{\omega}}_1{}'\tilde{\mathbf{Z}}_{it}, \qquad V_{it}^{P} = \mathbf{1}\{P_{it} \geq p_{0.75}\}, \qquad (3)$$

where $\hat{\boldsymbol{\omega}}_1$ is the leading eigenvector of the development-sample covariance matrix of $\tilde{\mathbf{Z}}_{it}$, $p_{0.75} = Q_{0.75}(P \mid \mathcal{D}_0)$, and its sign is normalized to yield a positive development-sample association with B_{it}. The PCA measure is interpreted as a latent statistical component rather than as a structural vulnerability factor. Hence $m \in \{B, S, P\}$ indexes alternative mappings $V_{it}^{m} = h_m(\mathbf{Z}_{it}; \boldsymbol{\theta}_m)$ of the same eight-dimensional information set.

For each definition m, the forecasting object is the one-step-ahead conditional probability of vulnerability. With $\mathcal{F}_{i,t-1}$ denoting the information available for country i through $t-1$,

$$V_{it}^{m} \mid \mathcal{F}_{i,t-1} \sim \text{Bernoulli}(p_{it}^{m}), \qquad p_{it}^{m} = E\left(V_{it}^{m} \mid \mathcal{F}_{i,t-1}\right) = f_m\left(\mathbf{X}_{i,t-1}^{D}, \mathbf{X}_{i,t-1}^{E}\right). \quad (4)$$

The conditioning set comprises domestic macro-fiscal variables,

$$\mathbf{X}_{i,t-1}^{D} = (Debt,\ Balance,\ Revenue,\ Expenditure,\ Growth,\ Inflation)'_{i,t-1},$$

and external variables,

$$\mathbf{X}^{E}_{i,t-1} = (CurrentAccount,\ FXDepreciation,\ REERChange,\ Exports,\ Imports)'_{i,t-1}.$$

All regressors are dated $t-1$, with changes and lags constructed only across adjacent calendar years. The conventional parametric benchmark restricts the conditional mean to $p^{m,L}_{it} = \Lambda(\alpha_m + \mathbf{X}_{i,t-1}'\boldsymbol{\beta}_m)$, whereas the candidate model class allows $f_m(\cdot)$ to be nonlinear and non-additive. The exercise is predictive throughout and imposes no structural or causal interpretation on $f_m(\cdot)$.

The sample chronology is imposed to identify genuine out-of-sample forecast performance. The raw panel begins in 1980, but the combination of one-year-ahead prediction and lagged annual transformations makes 1982 the first admissible forecast year. Using the entire panel jointly for model construction and evaluation would allow future observations to influence the forecasting rule against which they are subsequently assessed. The initial temporal design is therefore

$$\underbrace{1982-2005}_{\mathcal{D}\text{: development and model selection}} \qquad \underbrace{2006-2009}_{\mathcal{C}\text{: calibration}} \qquad \underbrace{2010-2017}_{\mathcal{E}^{I}\text{: held-out evaluation}} \qquad \underbrace{2018-2023}_{\mathcal{E}^{X}\text{: external evaluation}}.$$

Within $\mathcal{D}$, specification selection is conducted by expanding-window rolling-origin validation. For fold f, the historical estimation set $\mathcal{T}_f$ strictly precedes the validation set $\mathcal{V}_f$,

$$\max t(\mathcal{T}_f) < \min t(\mathcal{V}_f).$$

Because V^{B}_{it} is itself generated from estimated quantiles, outcome construction is nested within the recursive validation scheme. Specifically,

$$c_{k,f} = Q_{\tau_k}(Z_k \mid \mathcal{T}_f), \qquad V^{B}_{it,f} = \mathbf{1}\left\{\sum_{k=1}^{8} \mathbf{1}\left[d_k(Z_{kit} - c_{k,f}) \geq 0\right] \geq 3\right\}, \qquad (5)$$

with $\tau_k = 0.75$ for upper-tail and 0.25 for lower-tail risks. Thus, the validation block does not enter the parameters defining its own target. The same recursive restriction applies to preprocessing: the 1st -99th percentile clipping bounds, median-imputation parameters, and scaling parameters where required are estimated on $\mathcal{T}_f$ and carried unchanged into $\mathcal{V}_f$.

Let $\mathcal{M}$ denote the pre-specified library of 20 candidate specifications from 13 model families. The selected specification satisfies

$$\hat{r} = *\,arg\,lexmax_{r\in\mathcal{M}} \left\{ \overline{\mathrm{AUC}}_r,\ \overline{\mathrm{AP}}_r,\ \min_f \mathrm{AUC}_{rf},\ -\overline{\mathrm{BS}}_r \right\}, \qquad (6)$$

so that mean ROC-AUC is the primary criterion, followed sequentially by mean average precision, minimum fold-specific ROC-AUC, and mean Brier loss. Once selected on the baseline outcome, the model family and hyperparameters are held fixed. The same forecasting architecture is re-estimated under V^B, V^S, and V^P; it is not re-selected separately for each definition.

Probability calibration is separated from model estimation. If $\tilde{p}_{it}^m$ denotes the raw probability generated by the fitted model, the reported forecast is obtained from the logistic recalibration

$$\hat{p}_{it}^m = \Lambda[a_m^c + b_m^c \,\mathrm{logit}\,(\tilde{p}_{it}^m)]. \qquad (7)$$

For the internal exercise, estimation uses 1982-2005, calibration uses 2006-2009, and evaluation is confined to 2010-2017. For the principal external exercise, the locked specification is re-estimated using the information subsequently available through 2013, recalibrated on 2014-2017, and evaluated over 2018-2023:

$$\mathcal{T}^X = \{t \le 2013\}, \qquad \mathcal{C}^X = \{2014 \le t \le 2017\}, \qquad \mathcal{E}^X = \{2018 \le t \le 2023\},$$

with

$$\max t(\mathcal{T}^X) < \min t(\mathcal{C}^X) < \min t(\mathcal{E}^X), \qquad \mathcal{E}^X \cap (\mathcal{T}^X \cup \mathcal{C}^X) = \emptyset. \qquad (8)$$

This recursive re-estimation uses the historical information that would have been available before the external forecast period while preserving 2018-2023 as an untouched evaluation block. The parameters defining the vulnerability outcomes remain fixed at their development-sample estimates. The 2024 observations are reserved for a separate sensitivity exercise.

Persistence provides the principal economic benchmark. Let C_{it}^m denote the continuous vulnerability score underlying definition m, such that $C_{it}^B = B_{it}$, $C_{it}^S = S_{it}$, and $C_{it}^P = P_{it}$. The fitted persistence forecast is

$$q_{it}^{m} = \Pr\left(V_{it}^{m} = 1 \mid C_{i,t-1}^{m}\right) = \Lambda\left(a_{m}^{p} + b_{m}^{p} C_{i,t-1}^{m}\right). \quad (9)$$

Accordingly, the substantive comparison is between $f_m\left(\mathbf{X}_{i,t-1}\right)$ and the information contained in $C_{i,t-1}^{m}$ alone. Forecasts are compared on matched country-year observations using ROC-AUC and average precision for discrimination and the Brier score for probability accuracy. Incremental probability skill relative to fitted persistence is

$$\mathrm{BSS}_m = 1 - \frac{\mathrm{BS}_m(\hat{p})}{\mathrm{BS}_m(\hat{q})}, \qquad \mathrm{BS}_m(\hat{p}) = \frac{1}{N_m} \sum_{(i,t)\in\mathcal{E}_m} (V_{it}^{m} - \hat{p}_{it}^{m})^2. \quad (10)$$

Hence $\mathrm{BSS}_m > 0$ denotes an out-of-sample reduction in quadratic probability loss relative to fitted persistence. Sampling uncertainty is obtained from 1,000 country-clustered bootstrap replications. The same external forecasts are additionally evaluated conditional on $V_{i,t-1}^{m} = 0$ and $V_{i,t-1}^{m} = 1$ to distinguish vulnerability onset from continuation, while year-specific Spearman correlations, Kendall's τ, and upper-quintile overlap assess the stability of country-level surveillance rankings across the three outcome definitions. Detailed procedures for constructing the vulnerability measures are provided in Supplementary Appendix A, Section A.3. The forecasting design, recursive validation folds, candidate specifications, preprocessing, probability calibration, benchmarks, and uncertainty estimation are documented in Supplementary Appendix B (Online Resource 2), Sections B.1–B.5.

3. Empirical Results and Discussions

The empirical analysis examines how much one-year-ahead predictive information about fiscal vulnerability remains beyond persistence and how sensitive that incremental information and the resulting country-level surveillance priorities are to the definition of vulnerability. The results cover measurement, temporal forecast performance, incremental predictive content, vulnerability onset, and cross-definition ranking stability. Supplementary Appendix B reports additional analyses of alternative vulnerability thresholds and scaling rules, predictor availability, operating thresholds, country-level heterogeneity, and the 2024 post-evaluation sensitivity exercise (Sections B.6-B.10).

3.1 Vulnerability incidence and historical consistency

The analytical panel comprises 4,527 country-year observations for 127 developing economies over 1982-2024. Under the baseline breadth definition, which requires at least three of eight fiscal, macroeconomic, and external dimensions to exceed their adverse historical thresholds, 27.9% of country-years are classified as vulnerable. The corresponding incidence is 23.6% under the severity measure and 29.5% under the PCA-based statistical component. These differences indicate that vulnerability classifications are sensitive to the aggregation of the same underlying economic indicators.

Table 1 reports substantial variation across income groups. Breadth vulnerability declines from 38.2% of observations in low-income economies to 26.7% in lower-middle-income economies and 24.4% in upper-middle-income economies. The mean breadth score follows the same pattern, declining from 2.20 to 1.60 adverse dimensions. Severity-based vulnerability also decreases across the three groups, from 28.9% to 20.8%. While these differences are descriptive and reflect the sample and measurement rules employed, they are consistent with the importance of economic and fiscal fundamentals in assessments of debt distress (Kraay and Nehru, 2006).

Table 1: Incidence of fiscal vulnerability by income group

Income group	Economies	Country-years	Breadth vulnerable (%)	Severity vulnerable (%)	PCA vulnerable (%)	Mean breadth score	High debt (%)	Large deficit (%)	Low revenue (%)
Low income	24	819	38.2	28.9	18.4	2.20	32.5	26.0	35.3
Lower-middle income	50	1,846	26.7	24.0	26.5	1.79	20.7	26.5	24.2
Upper-middle income	53	1,862	24.4	20.8	37.4	1.60	10.7	17.6	8.4
Total	**127**	**4,527**	**27.9**	**23.6**	**29.5**	**1.79**	**18.7**	**22.8**	**19.7**

Notes: *Country-years contain complete observations for the eight dimensions entering the vulnerability measures. Breadth vulnerability equals one when at least three dimensions are in their adverse region. Income groups correspond to the static classifications in the June 2026 GMD release and are sample descriptors rather than historical annual classifications.*

The PCA-based measure displays a contrasting income-group pattern, with vulnerability increasing from 18.4% in low-income economies to 37.4% in upper-middle-income economies. Unlike breadth and severity, which explicitly identify adverse conditions, the PCA-based measure captures the dominant statistical variation across the same indicators and does not necessarily preserve their common adverse direction. Consequently, the observed reversal does not imply that

upper-middle-income economies face greater underlying fiscal distress. Rather, it demonstrates how the aggregation rule can alter the distribution and economic interpretation of the constructed outcome.

The component frequencies reinforce the multidimensional character of the breadth measure. High government debt is recorded in 18.7% of country-years and large fiscal deficits in 22.8%, alongside adverse expenditure, growth, current-account, inflation, and exchange-rate conditions documented in Supplementary Appendix A, Table A3.3 and Figure A3.2. These findings indicate that the breadth classification is not determined by public debt alone. This interpretation is consistent with evidence on the importance of debt composition for sovereign distress and external liabilities and current-account conditions for external-crisis risk (Catão and Milesi-Ferretti, 2014; Jarmuzek and Vesperoni, 2018).

Table 2 examines the historical consistency of the baseline breadth measure with independently recorded banking, sovereign-debt, and currency-crisis episodes. The crisis-indicator definitions, source chronologies, and country-year matching procedures are documented in Supplementary Appendix A, Section A.4.1 and Table A4.3, while the corresponding historical comparisons are reported in Table A4.4. Breadth vulnerability is observed in 56.4% of banking-crisis observations, 46.7% of sovereign-debt-crisis observations, and 78.2% of currency-crisis observations, compared with 28.2%, 28.4%, and 26.5%, respectively, outside the corresponding crisis episodes.

Table 2: Fiscal vulnerability and historical crisis episodes

Crisis type	Available country-years	Crisis observations	Economies with crisis	Vulnerable during crisis (%)	Vulnerable outside crisis (%)	Difference (p.p.)
Banking crisis	3,302	78	60	56.4	28.2	28.2
Sovereign-debt crisis	3,309	75	41	46.7	28.4	18.3
Currency crisis	3,325	147	85	78.2	26.5	51.7

Notes: Available country-years are complete-index observations for which the corresponding crisis indicator is observed. Differences equal the vulnerable share during crisis episodes minus the vulnerable share outside crisis periods. Crisis indicators are used solely as a historical-consistency diagnostic and are not forecasting targets. Crisis-indicator definitions, source chronologies, country-year coverage, and matching procedures are documented in Supplementary Appendix A, Section A.4.1 and Tables A4.3-A4.4.

The association is strongest for currency crises, for which vulnerability incidence is 51.7 percentage points higher during recorded crisis episodes, followed by banking crises (28.2 percentage points) and sovereign-debt crises (18.3 percentage points). These findings are

consistent with the interdependence of banking and currency disturbances and the importance of credit and external conditions in financial crises (Kaminsky and Reinhart, 1999; Gourinchas and Obstfeld, 2012). The sovereign-debt association is also consistent with evidence on the roles of macroeconomic fundamentals, international financial conditions, and sovereign-bank risk transmission (Arghyrou and Kontonikas, 2012; Kaminsky and Vega-García, 2016; Beqiraj et al., 2021). These associations should not be interpreted as evidence that the constructed vulnerability measure predicts realised crises. The comparison is contemporaneous, and the vulnerability and crisis indicators represent different empirical outcomes. The results provide a historical-consistency diagnostic rather than independent validation of the forecasting target.

Figure 1 presents the annual incidence of the three measures over 1982-2023. Breadth vulnerability is highest in 1982, at 61.8%, and rises sharply from 18.0% in 2019 to 48.3% in 2020 before declining to 25.6% in 2021. Severity-based vulnerability reaches 58.3% in 2020, while the PCA-based measure follows a different subsequent trajectory, remaining at 41.3% in 2021, 37.8% in 2022, and 35.1% in 2023.

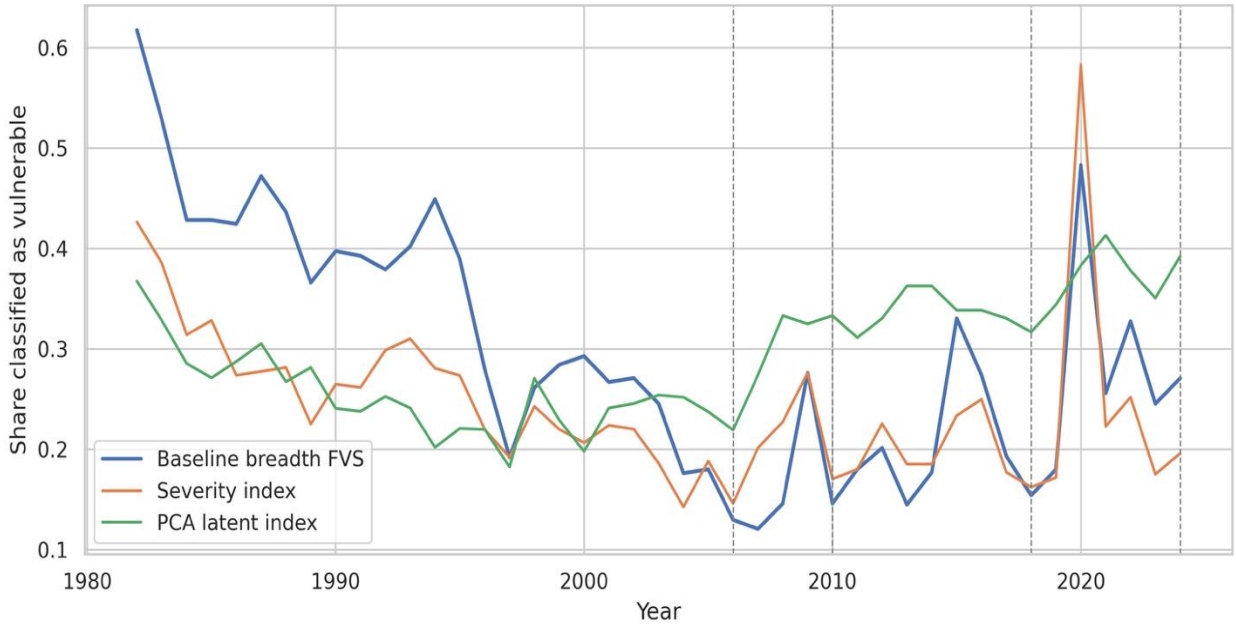


Figure 1: Annual prevalence of alternative fiscal-vulnerability measures, 1982-2023

Note: Annual breadth-vulnerability prevalence and country-year coverage are reported in Supplementary Appendix A, Table A4.2 and Figure A4.4.

The cross-sectional and temporal patterns establish that alternative constructions of fiscal vulnerability are related but not interchangeable. Their different distributions motivate the

subsequent analysis of whether forecast performance, incremental predictive information beyond persistence, and country-level surveillance rankings remain robust to the definition of vulnerability.

3.2 Model selection and temporal generalization

Model selection is conducted within the 1982-2005 development period using expanding-window rolling-origin validation, with vulnerability outcomes and predictor preprocessing reconstructed from the corresponding historical training observations. Table 3 reports the selected specification, its principal nonlinear competitors, and the conventional logistic benchmark. The complete library of 20 candidate specifications from 13 model families, together with the recursive validation design and selection results, is documented in Supplementary Appendix B, Sections B.1.1-B.2 and Tables B1.2, B2.1-B2.2.

The selection criterion identifies the depth-7 random forest, with a mean validation AUC of 0.876. Its advantage over XGBoost is marginal, however, at only 0.001 AUC points, while Extra Trees and Gradient Boosting achieve AUCs of 0.870. XGBoost also records slightly higher average precision and a lower Brier score than the selected specification. Although the nonlinear ensemble models outperform logistic regression in development-period discrimination, their relatively small differences indicate a competitive group of forecasting specifications rather than the clear superiority of a single algorithm. The random forest is therefore retained as a fixed forecasting specification for subsequent evaluation, without further model selection using the held-out samples.

Table 3: Rolling-origin model selection

Specification	Mean AUC	Worst-fold AUC	Mean AP	Mean Brier
Random Forest, depth 7	0.876	0.852	0.685	0.119
XGBoost, depth 2	0.875	0.854	0.686	0.108
Extra Trees, depth 6	0.870	0.857	0.684	0.172
Gradient Boosting, depth 2	0.870	0.849	0.683	0.111
Neural network, 16 units	0.861	0.843	0.660	0.118
Logistic regression, $C = 0.2$	0.829	0.824	0.652	0.148

Notes: Statistics are calculated across three expanding-window validation folds. Specification selection follows a lexicographic criterion based on mean ROC-AUC, mean average precision, worst-fold ROC-AUC, and mean Brier score. Outcome-construction parameters and predictor transformations are estimated separately within each training fold. The complete candidate library, fixed hyperparameters, and recursive model-selection results are reported in Supplementary Appendix B, Tables B2.1-B2.2.

Table 4 reports predictive performance in the 2010-2017 internal evaluation and the 2018-2023 external temporal evaluation. For the latter, the selected architecture is re-estimated using observations available through 2013, with probability recalibration conducted over 2014-2017. Model selection remains unchanged, and observations from the external evaluation period are excluded from estimation and recalibration. This chronological design provides a more demanding assessment of temporal generalisation than random country-year partitions, which can obscure the distinction between historical information and subsequent outcomes in persistent panel data (Hewamalage et al., 2023).

Table 4: Temporal predictive performance of the locked specification

Evaluation period	Observations	Events	AUC	AP	Brier	ECE_{10}
Internal, 2010–2017	989	204	0.838	0.595	0.120	0.039
External, 2018–2023	719	197	0.819	0.682	0.141	0.057

Notes: Entries report performance of the calibrated depth-7 random forest under the baseline breadth definition. Country-clustered 95% confidence intervals are 0.80-0.872 and 0.772-0.861 for internal and external AUC, respectively; 0.503-0.681 and 0.576-0.771 for average precision; and 0.101-0.141 and 0.119-0.161 for the Brier score. ECE_{10} denotes expected calibration error calculated using ten probability bins. Confidence intervals are conditional on the fitted forecasting pipeline and do not account for model-selection or vulnerability-measurement uncertainty.

The selected model retains substantial discrimination across both evaluation periods, with AUC declining moderately from 0.838 internally to 0.819 externally. Average precision increases from 0.595 to 0.682, although this partly reflects the higher vulnerability prevalence in the external sample (27.4% compared with 20.6%) and should not be interpreted as an unqualified improvement in predictive performance. Probability accuracy is weaker in the later period, with the Brier score increasing from 0.120 to 0.141 and expected calibration error from 0.039 to 0.057. These differences indicate some deterioration in probability forecasting under the later macroeconomic conditions, despite relatively stable ranking performance. However, the reported confidence intervals do not establish a statistically significant difference in performance between the two evaluation periods.

Figure 2 presents the corresponding ROC, precision–recall, and calibration diagnostics. The ROC and precision–recall curves indicate substantial predictive discrimination in both samples, with average precision exceeding the corresponding event-prevalence benchmarks. The calibration plots show broadly similar relationships between predicted probabilities and observed vulnerability frequencies, although deviations are more pronounced in parts of the external probability distribution. Together, these diagnostics suggest that the model retains useful ranking

information over time, while the accuracy of its predicted probabilities is more sensitive to the evaluation period.

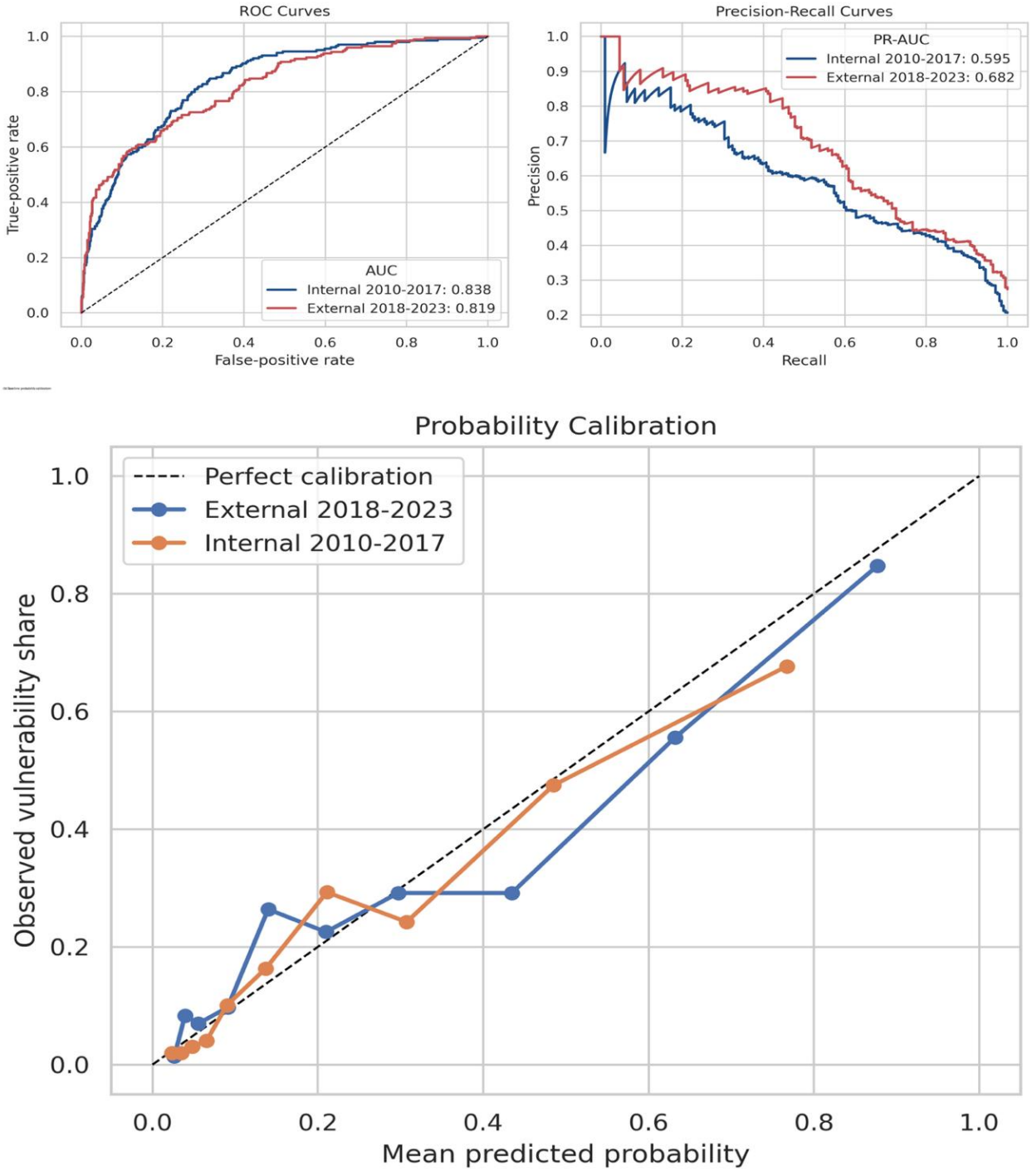


Figure 2: Discrimination and calibration under temporal evaluation

At the prespecified probability threshold of 0.30, the external forecasts achieve recall of 0.675, precision of 0.538, and a false-alarm rate of 0.218. These operating characteristics illustrate the trade-off between detecting vulnerable observations and generating false warnings rather than identifying an optimal surveillance threshold. The appropriate threshold depends on the relative institutional costs of missed vulnerabilities and false alarms; threshold-dependent operating characteristics and sensitivity analyses are reported in Supplementary Appendix B, Section B.8.

The results demonstrate that the selected model retains substantial predictive performance beyond the development period, although probability accuracy weakens in the later evaluation. Strong out-of-sample performance does not, however, establish that the multivariate information set provides meaningful gains beyond the persistence of vulnerability itself. The next section addresses this distinction through comparisons with fitted persistence and conventional forecasting benchmarks.

3.3 Forecastability and incremental information beyond persistence

Table 5 presents the central forecasting comparison for the 2018-2023 external evaluation. For each vulnerability definition, the selected multivariate model is assessed against fitted lagged-score persistence on matched country-year observations, with logistic regression included as an additional benchmark. The comparison distinguishes the overall predictability of fiscal vulnerability from the incremental predictive information provided by the broader macro-fiscal predictor set.

Table 5: External forecasting performance across core vulnerability definitions

Vulnerability definition	Selected AUC	Logit AUC	Persistence AUC	Δ AUC vs persistence, 95% CI	Selected AP	Persistence AP	Δ AP vs persistence, 95% CI	Brier skill vs persistence
Breadth	0.819	0.799	0.797	+0.022 [0.006, 0.039]	0.682	0.603	+0.079 [0.050, 0.109]	+5.7%
Severity	0.813	0.761	0.793	+0.020 [0.004, 0.038]	0.706	0.618	+0.088 [0.044, 0.130]	+16.2%
PCA statistical component	0.959	0.964	0.963	−0.004 [−0.012, 0.004]	0.931	0.939	−0.008 [−0.025, 0.007]	−1.8%

Notes: Results refer to the 2018–2023 external evaluation. Selected-model performance statistics are reported for 719 country-year observations, while paired comparisons with fitted lagged-score persistence use 718 matched observations from 124 economies. Confidence intervals are based on 1,000 paired country-clustered bootstrap replications and are conditional on the fitted forecasting pipelines. Brier skill is defined as $1 - \mathrm{BS}_{\mathrm{selected}}/\mathrm{BS}_{\mathrm{persistence}}$*; positive values indicate lower probability loss for the selected model. The persistence benchmarks and country-clustered paired comparisons are documented in Supplementary Appendix B, Sections B.4–B.5, particularly Tables B4.1 and B5.2.*

Under the baseline breadth definition, the selected model achieves an AUC of 0.819 compared with 0.797 for fitted persistence. The paired improvement of 0.022 is statistically distinguishable from zero, while average precision rises from 0.603 to 0.682. The Brier score also declines from approximately 0.150 to 0.141, corresponding to a 5.7% reduction in probability loss. Although these gains are moderate, they indicate that the broader macro-fiscal information set provides predictive content not fully captured by the previous continuous breadth score.

The severity definition yields a similar improvement in discrimination but a larger gain in probability accuracy. The selected model records an AUC of 0.813 against 0.793 for fitted persistence, while average precision increases from 0.618 to 0.706. Its Brier score declines from approximately 0.157 to 0.132, representing a 16.2% improvement relative to persistence. Thus, although breadth and severity exhibit comparable incremental discrimination, the severity construction yields a larger reduction in probability loss. This difference suggests that the forecasting value of additional macro-fiscal information depends not only on whether adverse conditions are present but also on how their intensity enters the vulnerability measure.

The PCA-based construction produces a contrasting result. Absolute discrimination is exceptionally high across all three forecasting specifications, with AUCs ranging from 0.959 to 0.964. However, the selected model does not outperform fitted persistence: its paired AUC difference is −0.004, average precision is slightly lower, and Brier skill is −1.8%. Neither the AUC nor the average-precision difference is statistically distinguishable from zero. Logistic regression also achieves a marginally higher AUC than the selected nonlinear specification. These results indicate that the previous PCA score captures most of the information relevant for forecasting the subsequent PCA-based classification, leaving little detectable incremental benefit from the broader predictor set. The high absolute performance of this measure should not, however, be interpreted as evidence that it provides a superior economic representation of fiscal vulnerability.

Figure 3 illustrates the distinction between absolute predictability and incremental forecasting performance. The selected model improves upon fitted persistence under breadth and severity, although the discrimination gains are relatively small. Under PCA, all three specifications exhibit very high and closely aligned AUCs. The comparison therefore demonstrates that strong predictive discrimination does not necessarily imply that a multivariate forecasting model contributes substantial information beyond the previous vulnerability state.

The sensitivity results provide further evidence of measurement dependence. Relative to fitted persistence, Brier skill remains positive under robust-scaled severity (8.2%), alternative component thresholds of 20/80 (9.5%) and 30/70 (5.0%), and breadth definitions requiring at least two (3.6%) or four (2.2%) adverse dimensions. Leave-one-component-out specifications attain external AUCs between 0.795 and 0.861, while the complete-predictor sample produces an AUC of 0.822, close to the baseline estimate. These exercises indicate that predictive discrimination is retained under several alternative constructions and sample restrictions, although the magnitude of incremental probability gains remains sensitive to the measurement rule. Detailed robustness results for alternative outcome definitions, component exclusions, and predictor-completeness restrictions are reported in Supplementary Appendix B, Section B.6 and Tables B6.1-B6.3.

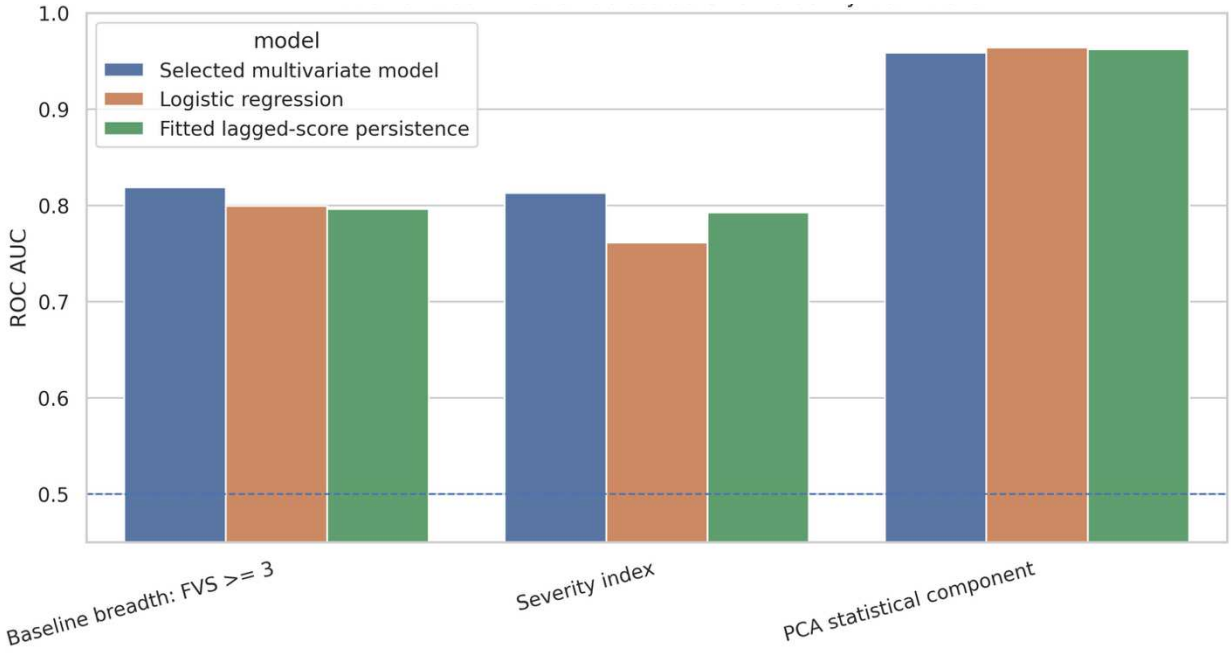


Figure 3: External discrimination across core vulnerability definitions

The findings offer a more qualified interpretation of machine-learning performance in fiscal-vulnerability forecasting. Previous research has shown that greater model flexibility does not necessarily produce superior out-of-sample crisis predictions (Beutel et al., 2019). The present evidence is consistent with this qualification: the selected model records higher discrimination than logistic regression under breadth and severity, but its breadth advantage is not statistically distinguishable from zero, and the ranking reverses under PCA. More importantly, the sensitivity of incremental gains to outcome construction reinforces the relevance of multidimensional fiscal-

vulnerability assessment (Stoian et al., 2018). The principal finding is therefore not that nonlinear methods consistently outperform conventional approaches, but that the value of multivariate forecasting beyond persistence depends on how fiscal vulnerability is defined. This distinction motivates the subsequent examination of new vulnerability onsets and country-level surveillance rankings.

3.4 Vulnerability onset versus continuation

Aggregate forecasting performance may partly reflect the persistence of existing vulnerability rather than the ability to identify newly emerging adverse conditions. To distinguish these cases, the 2018-2023 external evaluation is partitioned according to the preceding year's vulnerability state. For each definition m, onset observations satisfy $V_{i,t-1}^{(m)} = 0$, while continuation observations satisfy $V_{i,t-1}^{(m)} = 1$. The same externally generated probabilities are evaluated in both subsamples; no transition-specific forecasting model is estimated. This distinction follows the early-warning literature, which recognises that combining crisis onsets with continuing or post-crisis observations can obscure the predictive performance of warning systems (Bussière and Fratzscher, 2006; Dawood et al., 2017).

Table 6: External forecast performance for vulnerability onset and continuation

Vulnerability definition	Onset observations (events)	Onset AUC	ΔAUC vs persistence, 95% CI	Onset AP	ΔAP vs persistence, 95% CI	Continuation AUC
Breadth	528 (84)	0.721	+0.043 [0.010, 0.078]	0.338	+0.098 [0.045, 0.169]	0.779
Severity	533 (76)	0.663	+0.005 [−0.021, 0.036]	0.275	+0.003 [−0.065, 0.067]	0.829
PCA statistical component	458 (37)	0.875	−0.000 [−0.032, 0.037]	0.393	+0.057 [−0.065, 0.180]	0.814

Notes: Onset observations are country-years classified as non-vulnerable in the preceding year; events indicate transitions into vulnerability in the current year. Continuation observations are country-years classified as vulnerable in the preceding year, within which forecasts distinguish continued vulnerability from exits. AUC and AP are calculated from the selected model's external predictions. Paired differences relative to fitted lagged-score persistence are evaluated on matched onset observations, with 95% confidence intervals obtained through country-clustered bootstrap resampling. Onset and continuation AUCs describe different subsamples and are not formal paired comparisons. Transition-specific forecasting results and paired comparisons with fitted persistence are reported in Supplementary Appendix B, Section B.7 and Tables B7.1-B7.2.

The breadth measure provides the clearest evidence of incremental early-warning information (Refer Table 6). Among previously non-vulnerable country-years, the selected model achieves an

onset AUC of 0.721 and an average precision of 0.338. Relative to fitted persistence, discrimination improves by 0.043 AUC points and average precision by 0.098, with both confidence intervals excluding zero. These findings indicate that the full-sample forecasting gains reported in Section 3.3 are not attributable solely to the continuation of existing vulnerability. The broader macro-fiscal predictor set also contains information relevant to identifying new breadth-based vulnerability episodes.

The results differ under the severity measure. Although onset remains moderately predictable, with an AUC of 0.663, its incremental AUC and average-precision gains over persistence are small and statistically indistinguishable from zero. Discrimination is higher among previously vulnerable observations, where continuation AUC reaches 0.829. This pattern suggests that the strong aggregate forecasting performance under severity may be more closely associated with distinguishing the persistence or resolution of existing vulnerability than with generating additional information about new onsets. However, the difference between onset and continuation AUCs is descriptive and does not establish a statistically significant difference between the two forecasting tasks.

The PCA-based measure again exhibits high absolute predictability without corresponding incremental gains. Its onset AUC reaches 0.875, but the difference relative to fitted persistence is effectively zero, while the average-precision improvement is imprecisely estimated. The continuation AUC is 0.814. Consistent with Section 3.3, these findings indicate that much of the predictive information for the PCA-based outcome is already contained in its preceding continuous score. High onset discrimination should therefore not be interpreted as evidence that the broader predictor set generates additional early-warning information. The distinction between onset and continuation is consistent with earlier research demonstrating that the treatment of crisis and post-crisis observations affects the evaluation of early-warning models (Bussière and Fratzscher, 2006; Dawood et al., 2017; Hamdaoui, 2016). The present findings extend this distinction to constructed measures of fiscal vulnerability. While persistence remains relevant for monitoring prolonged macro-fiscal weakness, incremental information about transitions into vulnerability is more limited and varies across definitions. In particular, the breadth measure provides statistically distinguishable gains for new vulnerability onsets, whereas corresponding gains are not established for severity or the PCA-based measure. The early-warning value of multivariate forecasting

therefore depends not only on model performance but also on how vulnerability is defined and whether surveillance prioritises emerging or continuing fiscal pressures.

3.5 Stability of country-level surveillance forecasts

Aggregate forecasting performance does not necessarily imply agreement over which economies are identified as relatively vulnerable. For fiscal surveillance, the stability of country-level rankings is therefore important, particularly when forecasts are used to prioritise analytical attention. Figure 4 and Table 7 examine the agreement between annual country rankings generated by the breadth, severity, and PCA-based vulnerability forecasts over 2018-2023.

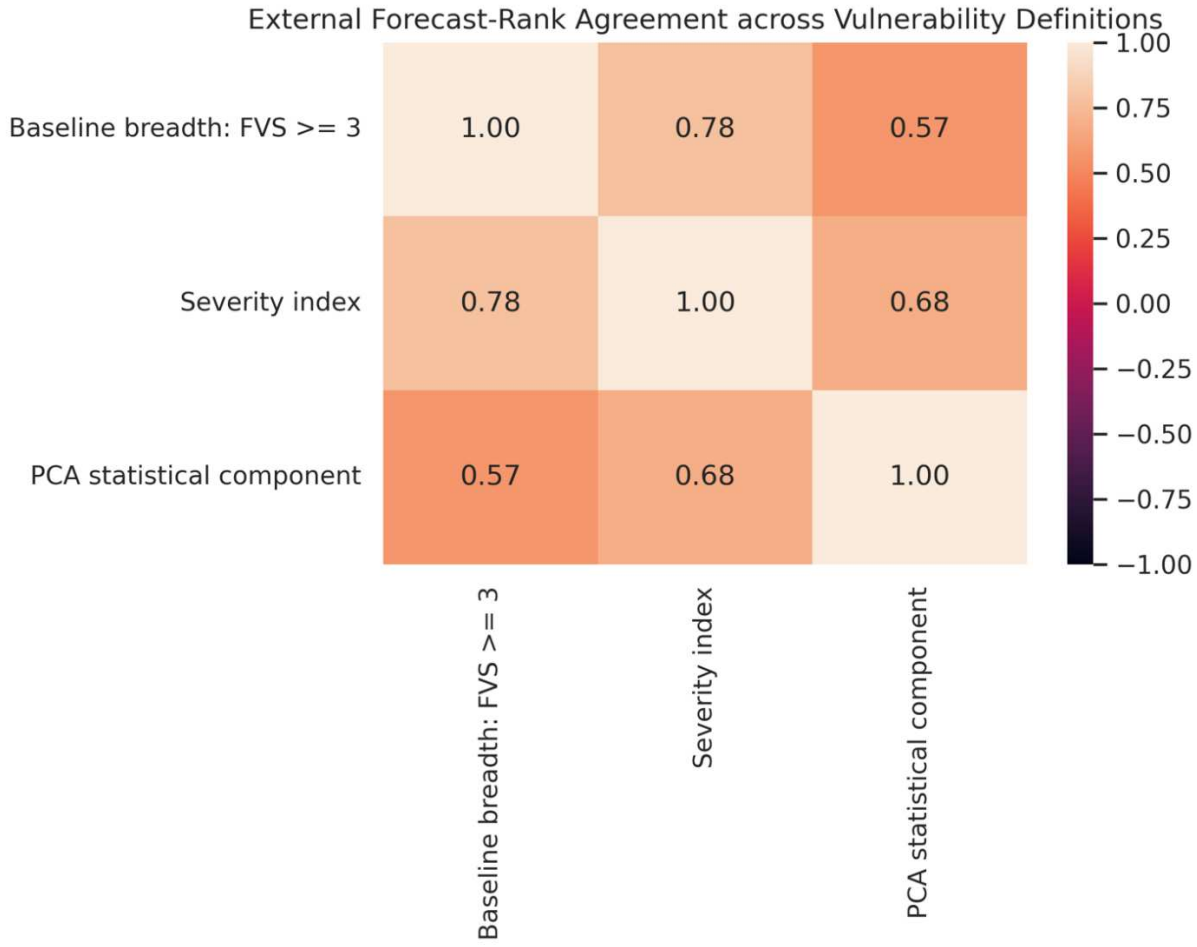


Figure 4: Cross-definition agreement in country-level surveillance forecasts

Figure 4 summarises average cross-definition ranking agreement, showing the strongest association between breadth and severity and the weakest between breadth and PCA. The annual correlation ranges reported in Table 7 indicate that this ordering is maintained throughout the

2018-2023 external-evaluation period. Breadth and severity produce the most similar country rankings, with mean Spearman and Kendall correlations of 0.775 and 0.578, respectively. Nevertheless, their average top-quintile overlap is only 60.3%, indicating that approximately two-fifths of the economies selected for the highest surveillance-priority group differ between the two definitions. Agreement is substantially weaker between breadth and PCA, with a mean Spearman correlation of 0.566 and top-quintile overlap of 38.4%. Thus, even where forecasting performance is strong across definitions, the countries receiving relatively high predicted vulnerability can differ considerably.

The ranking differences have implications for the use of early-warning models in fiscal surveillance. Earlier research has emphasised the role of macroeconomic indicators in identifying emerging risks and supporting precautionary policy responses (Kaminsky et al., 1998; Bussière and Fratzscher, 2006). The present findings highlight an additional consideration: surveillance priorities may also depend on how the underlying vulnerability outcome is constructed. The stronger agreement between breadth and severity is consistent with their common emphasis on adverse macro-fiscal conditions, whereas the PCA-based construction reflects a different statistical aggregation of the same indicators. Consequently, aggregate predictive accuracy alone is insufficient to establish the stability of country-level surveillance rankings.

Table 7: Annual rank agreement across vulnerability definitions, 2018–2023

Forecast pair	Mean Spearman	Annual range	Mean Kendall τ	Mean top-20% overlap
Breadth – Severity	**0.775**	0.722–0.813	**0.578**	**0.603**
Breadth – PCA	0.566	0.514–0.603	0.387	0.384
Severity – PCA	0.684	0.649–0.721	0.497	0.510

Notes: Statistics are averaged across annual cross-country comparisons for 2018–2023. Spearman's rank correlation and Kendall's τ measure overall ranking agreement, while top-20% overlap measures the proportion of economies common to the two highest predicted-probability quintiles. Annual cross-definition ranking agreement and country-level forecast summaries are reported in Supplementary Appendix B, Section B.9 and Tables B9.1–B9.2.

Country-level forecasts nevertheless reveal some recurring signals across the alternative measures. The country-level results in Supplementary Appendix B, Table B9.2 show substantial variation in mean predicted vulnerability probabilities, with Yemen, Sudan, Lebanon, and Dominica among the economies receiving relatively high breadth-based forecasts during the external evaluation period. Their relative positions differ under severity and PCA, however, reinforcing the importance of measurement choice. These predictions should not be interpreted as sovereign-default probabilities or comprehensive country risk ratings. They describe the probability of meeting the

corresponding constructed vulnerability definition, conditional on the information included in the pooled forecasting model.

This distinction is important because country-specific fiscal sustainability also depends on institutional capacity, debt structure, financing arrangements, and other characteristics that may not be fully represented in the forecasting framework. The use of multiple indicators to identify macroeconomic weaknesses is well established in early-warning research (Kaminsky et al., 1998), but such indicators do not constitute structural estimates of sovereign solvency. Model-based vulnerability probabilities should therefore be interpreted as comparative surveillance signals requiring further country-specific assessment.

The results show that robustness in aggregate forecastability does not imply robustness in the cross-country allocation of surveillance attention. Breadth and severity generate relatively similar rankings, whereas the PCA-based measure produces a materially different ordering despite its high predictive accuracy. Together with the incremental forecasting results in Section 3.3, this establishes that the definition of fiscal vulnerability affects both the information gained beyond persistence and the identification of economies warranting closer monitoring.

3.6 Predictor Importance Across Vulnerability Definitions

Figure 5 reports out-of-sample permutation importance for the selected forecasting model under the three vulnerability definitions. Fiscal balance and government debt are the leading predictors under breadth and severity, while government debt records the highest importance under the PCA-based measure. Their prominence is consistent with fiscal-distress research identifying fiscal positions, indebtedness, and broader macroeconomic conditions as relevant indicators of fiscal vulnerability (Bruns and Poghosyan, 2018; Schaechter et al., 2014).

The relative importance of other macro-fiscal indicators varies across definitions. Government expenditure and current-account conditions contribute notably under breadth, while inflation is relatively more important under severity. Under PCA, government debt and fiscal balance remain prominent, alongside current-account conditions. These differences suggest that the construction of the vulnerability outcome influences which aspects of the macro-fiscal information set contribute most to predictive discrimination.

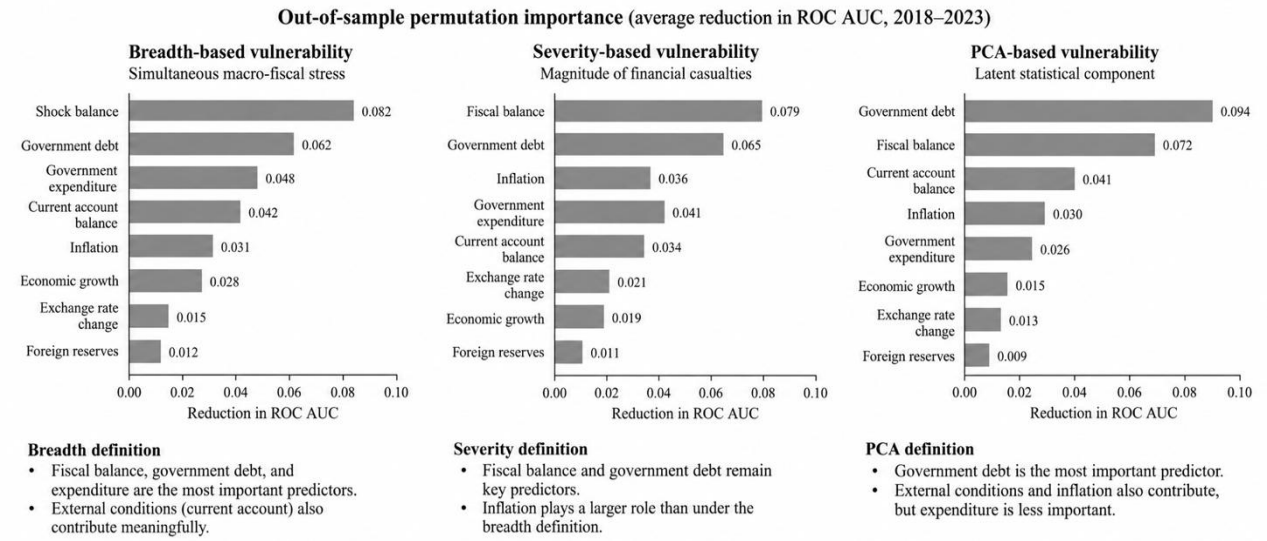


Figure 5: Key predictors of fiscal vulnerability across alternative definitions

Notes: Bars represent the reported reduction in ROC AUC following permutation of individual predictors during the 2018–2023 external evaluation. Higher values indicate greater predictive importance. Predictor contributions should not be interpreted causally. The permutation-importance methodology and additional predictor-importance diagnostics are presented in Supplementary Appendix B, Section B.8.4. Figure 5 compares the predictive importance of individual macro-fiscal indicators across the three vulnerability definitions using the 2018-2023 external-evaluation sample.

The findings reinforce the importance of multidimensional fiscal-risk assessment, while showing that predictor rankings are not invariant to the definition of vulnerability. They should nevertheless be interpreted cautiously: permutation importance measures the reduction in predictive performance when a variable is disrupted, not its independent economic effect. Correlation among fiscal and macroeconomic predictors may also affect the allocation of importance across variables. The results therefore identify indicators used by the forecasting model rather than structural determinants of fiscal vulnerability.

4. Conclusion, Policy Implications and Limitations

This study examines whether one-year-ahead fiscal-vulnerability forecasts provide predictive information beyond persistence and whether that information remains robust across alternative definitions of vulnerability. Using 127 developing economies and three measures constructed from the same eight fiscal, macroeconomic, and external dimensions, the analysis finds that fiscal

vulnerability is highly forecastable, but the incremental value of multivariate forecasting depends on how vulnerability is measured.

The distinction between absolute forecastability and incremental predictive content is central to the findings. Under the breadth and severity definitions, the selected multivariate model improves upon fitted lagged-score persistence, with relatively modest gains in discrimination and larger improvements in probability accuracy under severity. In contrast, the PCA-based statistical component exhibits exceptionally high predictability, but additional macro-fiscal predictors provide no detectable improvement over persistence. The onset analysis further shows that breadth-based forecasts contain incremental information about transitions into vulnerability, whereas corresponding gains are not statistically established under severity or PCA. These findings demonstrate that strong forecasting performance does not necessarily imply substantial predictive information beyond the previous vulnerability state.

The results have implications for fiscal surveillance. Although breadth and severity generate relatively similar country rankings, agreement with the PCA-based construction is considerably weaker, particularly among economies assigned to the highest predicted-vulnerability quintile. Consequently, the definition of vulnerability influences not only forecast performance but also the relative priority assigned to countries for further assessment. Fiscal-monitoring frameworks should therefore make their measurement assumptions explicit and examine whether surveillance conclusions remain stable across plausible alternative definitions. Agreement across measures may provide a useful basis for prioritising additional analysis, while divergent signals should prompt closer examination of the underlying fiscal and macroeconomic conditions rather than automatic reclassification of country risk. The findings also suggest distinguishing surveillance aimed at detecting emerging vulnerability from monitoring the persistence of existing pressures.

Several limitations qualify these conclusions. The three vulnerability measures are derived from a common indicator set, so the analysis assesses sensitivity to alternative measurement constructions rather than validity against an independent measure of fiscal distress. Historical crisis comparisons provide descriptive consistency checks, not evidence that the constructed outcomes anticipate realised crises. Moreover, the pooled forecasting framework cannot fully capture country-specific institutional arrangements, debt composition, financing conditions, or contingent liabilities. Predicted probabilities therefore represent the likelihood of meeting a particular constructed

vulnerability definition, not sovereign-default probabilities or structural assessments of fiscal sustainability.

Data availability and temporal coverage impose further constraints. The panel is unbalanced, income-group classifications are static, and incomplete macro-fiscal histories may affect the representativeness of the analytical sample. Although chronological evaluation reduces look-ahead bias and the 2018-2023 external period provides evidence of out-of-time performance, the results do not guarantee stability under future economic regimes. The 2024 observations are reserved for sensitivity analysis rather than the headline evaluation.

The findings establish that the definition of fiscal vulnerability is consequential for both forecasting and surveillance. Vulnerability may be highly predictable under alternative constructions, but the information gained beyond persistence, the identification of new vulnerability episodes, and the relative ranking of economies are not invariant to measurement choice. Robust fiscal-vulnerability forecasting therefore requires attention not only to model selection and predictive accuracy but also to the economic interpretation and stability of the outcome being forecast.

Data and Code Availability

The data used in this study are obtained from the June 2026 release of the Global Macro Database (GMD, version 2026_06), which is publicly available. The complete replication package, including processed datasets, Python scripts, and supporting materials used for data preparation, vulnerability measurement, model estimation, forecasting, robustness analyses, and generation of tables and figures, will be provided to the journal during the review process upon request. Upon acceptance for publication, the replication package will be deposited in a publicly accessible repository to ensure transparency, reproducibility, and independent verification of the empirical findings.

Reference List

Adger, W. N. (2006). Vulnerability. *Global Environmental Change*, 16(3), 268-281. doi:10.1016/j.gloenvcha.2006.02.006.

Anderson, C. C., Hagenlocher, M., Renaud, F. G., Cutter, S. C., & Emrich, C. T. (2019). Comparing index-based vulnerability assessments in the Mississippi Delta: Implications of contrasting theories, indicators, and aggregation methodologies. *International Journal of Disaster Risk Reduction*, 39. doi:10.1016/j.ijdrr.2019.101128.

Arghyrou, M. G., & Kontonikas, A. (2012). The EMU sovereign-debt crisis: Fundamentals, expectations and contagion. *Journal of International Financial Markets, Institutions and Money*, 22(4), 658-677. doi:10.1016/j.intfin.2012.03.003.

Beqiraj, E., Patella, V., & Tancioni, M. (2021). Fiscal stance and the sovereign risk pass-through. *Economic Modelling*, 102. doi:10.1016/j.econmod.2021.105573.

Beutel, J., List, S., & Schweinitz, G. V. (2019). Does machine learning help us predict banking crises? *Journal of Financial Stability*, 45. doi:10.1016/j.jfs.2019.100693.

Bi, H., Shen, W., & Yang, S. C. S. (2016). Fiscal limits in developing countries: A DSGE Approach. *Journal of Macroeconomics*, 49, 119-130. doi:10.1016/j.jmacro.2016.06.002.

Bluwstein, K., Buckmann, M., Joseph, A., Kapadia, S., & Şimşek, Ö. (2023). Credit growth, the yield curve and financial crisis prediction: Evidence from a machine learning approach. *Journal of International Economics*, 145. doi:10.1016/j.jinteco.2023.103773.

Bruns, M., & Poghosyan, T. (2018). Leading indicators of fiscal distress: evidence from extreme bounds analysis. *Applied Economics*, 50(13), 1454-1478. doi:10.1080/00036846.2017.1366639.

Bussière, M., & Fratzscher, M. (2006). Towards a new early warning system of financial crises. *Journal of International Money and Finance*, 25(6), 953-973. doi:10.1016/j.jimonfin.2006.07.007.

Catão, L. A. V., & Milesi-Ferretti, G. M. (2014). External liabilities and crises. *Journal of International Economics*, 94(1), 18-32. doi:10.1016/j.jinteco.2014.05.003.

Cipollini, A., & Kapetanios, G. (2009). Forecasting financial crises and contagion in Asia using dynamic factor analysis. *Journal of Empirical Finance*, 16(2), 188-200. doi:10.1016/j.jempfin.2008.10.004.

Dawood, M., Horsewood, N. J., & Strobel, F. (2017). Predicting sovereign debt crises: An Early Warning System approach. *Journal of Financial Stability*, 28, 16-28. doi:10.1016/j.jfs.2016.11.008.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, 22(3), 443-473. doi:10.1016/j.ijforecast.2006.01.001.

Du Plessis, E., & Fritsche, U. (2025). New forecasting methods for an old problem: Predicting 147 years of systemic financial crises. *Journal of Forecasting*, 44(1), 3-40. doi:10.1002/for.3184.

Fioramanti, M. (2008). Predicting sovereign debt crises using artificial neural networks: A comparative approach. *Journal of Financial Stability*, 4(2), 149-164. doi:10.1016/j.jfs.2008.01.001.

Fiordelisi, F., Girardone, C., Minnucci, F., & Ricci, O. (2020). On the nexus between sovereign risk and banking crises. *Journal of Corporate Finance*, 65. doi:10.1016/j.jcorpfin.2020.101717.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359-378. doi:10.1198/016214506000001437.

Gourinchas, P. O., & Obstfeld, M. (2012). Stories of the twentieth century for the twenty-first. *American Economic Journal: Macroeconomics*, 4(1), 226-265. doi:10.1257/mac.4.1.226.

Hamdaoui, M. (2016). Are systemic banking crises in developed and developing countries predictable? *Journal of Multinational Financial Management*, 37-38, 114-138. doi:10.1016/j.mulfin.2016.09.002.
Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices. *Data Mining and Knowledge Discovery*, 37(2), 788-832. doi:10.1007/s10618-022-00894-5.
Jarmuzek, M., & Vesperoni, E. R. (2018). The role of debt profile vulnerabilities in sovereign distress. *Applied Economics Letters*, 25(13), 928-935. doi:10.1080/13504851.2017.1386274.
Kaminsky, G. L., Lizondo, S., & Reinhart, C. M. (1998). Leading Indicators of Currency Crises. International Monetary Fund.
Kaminsky, G. L., & Reinhart, C. M. (1999). The twin crises: The causes of banking and balance-of-payments problems. *American Economic Review*, 89(3), 473-500. doi:10.1257/aer.89.3.473.
Kaminsky, G. L., & Vega-García, P. (2016). Systemic and idiosyncratic sovereign debt crises. *Journal of the European Economic Association*, 14(1), 80-114. doi:10.1111/jeea.12165.
Kraay, A. C., & Nehru, V. (2006). When is external debt sustainable? *World Bank Economic Review*, 20(3), 341-365. doi:10.1093/wber/lhl006.
Kuek, T. H., Chin-Hong, P., & Arip, M. A. B. (2020). Financial Vulnerability and Economic Dynamics in Malaysia. *Journal of Central Banking Theory and Practice*, 9(1), 55-73. doi:10.2478/jcbtp-2020-0023.
Liu, L., Chen, C., & Wang, B. (2022). Predicting financial crises with machine learning methods. *Journal of Forecasting*, 41(5), 871-910. doi:10.1002/for.2840.
Müller, K., Xu, C., Lehbib, M., & Chen, Z. (2025) 'The global macro database: A new international macroeconomic dataset' *NBER Working Paper*. National Bureau of Economic Research.
Rajaguru, G., Khan, S. U., & Ur Rahman, H. (2021). Analysis of Australia's Fiscal Vulnerability to Crisis. *Journal of Risk and Financial Management*, 14(7). doi:10.3390/jrfm14070297.
Schaechter, A., Alper, C. E., Arbatli-Saxegaard, E. C., Shabunina, A., & Weber, A. (2014). A toolkit to assess fiscal vulnerabilities and risks in advanced economies. *Applied Economics*, 46(6), 650-660. doi:10.1080/00036846.2013.861589.
Stoian, A. M., Obreja Brașoveanu, L., Braşoveanu, I. V., & Dumitrescu, B. A. (2018). A framework to assess fiscal vulnerability: Empirical evidence for European Union countries. *Sustainability (Switzerland)*, 10(7). doi:10.3390/su10072482.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1AppendixA.pdf
- ESM2AppendixB.pdf