Supplementary Appendix B

Forecasting Implementation, Benchmarks, and Robustness

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

This Supplementary Appendix B documents the predictive implementation underlying the empirical results in the main manuscript. It provides the complete forecasting protocol, including the chronological validation design, fold-specific outcome reconstruction and preprocessing, candidate-model comparison, probability calibration, persistence and conventional benchmarks, performance metrics, uncertainty estimation, robustness exercises, vulnerability-onset analysis, operating-threshold diagnostics, country-level surveillance forecasts, and post-evaluation sensitivity analysis.

The forecasting exercise begins from the analytical panel and the breadth, severity, and PCA-based vulnerability outcomes constructed in Supplementary Appendix A. For vulnerability definition $m \in \{B, S, P\}$, the one-year-ahead forecasting target is

$$p_{it}^{m} = \Pr\left(V_{it}^{m} = 1 \mid \mathcal{F}_{i,t-1}\right),$$

where V_{it}^{m} denotes the binary vulnerability state for economy $i$in year t, and $\mathcal{F}_{i,t-1}$ denotes the information available through year $t-1$. The forecasting information set consists of eleven lagged fiscal, domestic macroeconomic, and external-sector predictors. Accordingly, all reported forecasts are genuinely one-year-ahead predictions rather than contemporaneous classifications.

The empirical design preserves temporal ordering at every stage. Model selection is conducted using expanding-window rolling-origin validation within the 1982–2005 development period. Within each fold, vulnerability thresholds and other outcome-construction parameters are estimated exclusively from the corresponding training history before labels are assigned to validation observations. Predictor preprocessing is likewise fitted using training-fold observations only and transferred unchanged to the validation period. Probability calibration and final evaluation are conducted in subsequent non-overlapping periods. This nesting prevents future validation or evaluation observations from influencing outcome construction, preprocessing, model selection, estimation, or probability recalibration.

Candidate specifications are compared under a common forecasting information set and validation protocol. Model selection follows a prespecified lexicographic rule based sequentially on mean ROC AUC, mean average precision, worst-fold ROC AUC, and mean Brier score. The selected specification is a random forest with maximum depth seven. The candidate-model exercise is used to identify a competitive predictive specification; it is not treated as a source of methodological novelty.

Particular attention is given to persistence. Because fiscal vulnerability is serially dependent, strong out-of-sample performance may arise largely from information contained in the previous vulnerability state. The selected multivariate model is therefore evaluated against a fitted lagged-score persistence benchmark, conventional logistic regression, lagged binary-state persistence, and a historical-event-rate benchmark. Comparisons with fitted persistence are conducted on matched observations so that differences in discrimination and probability accuracy are not induced by changes in evaluation samples.

Forecast uncertainty is assessed using country-clustered resampling, preserving within-economy dependence across repeated annual observations. Robustness analysis then examines whether the forecasting conclusions survive alternative vulnerability constructions, breadth thresholds, robust severity scaling, predictor-completeness restrictions, and component-ablation exercises. Separate analyses distinguish transitions into vulnerability from continuation of existing vulnerability and assess whether country-level surveillance rankings remain stable across outcome definitions.

Country-level results are generated from pooled cross-country forecasting models. The reported quantities are therefore country-specific predicted vulnerability probabilities and associated surveillance rankings, not estimates from separate country-specific models, default probabilities, or country risk ratings. All results are predictive and are not interpreted causally.

The appendix is organized as follows. Section B.1 defines the forecasting estimand, temporal design, and leakage-control protocol. Section B.2 documents fold-specific preprocessing, candidate-model search, and model selection. Section B.3 describes probability calibration and forecast-performance metrics. Section B.4 specifies the persistence and other forecasting benchmarks. Section B.5 documents uncertainty estimation and paired model comparisons. Section B.6 reports robustness to alternative vulnerability definitions, scaling rules, and sample restrictions. Section B.7 examines vulnerability onset and continuation. Section B.8 reports

operating-threshold and supplementary forecasting diagnostics. Section B.9 presents country-level surveillance forecasts and cross-definition ranking agreement. Section B.10 reports the 2024 post-evaluation sensitivity exercise.

B.1 Forecasting estimand, temporal design, and leakage control

For economy i, year t, and vulnerability definition $m \in \{B, S, P\}$, the forecasting object is the one-year-ahead conditional probability

$$V_{it}^{m} \mid \mathcal{F}_{i,t-1} \sim \text{Bernoulli}\ (p_{it}^{m}),$$

with

$$p_{it}^{m} = \Pr\left(V_{it}^{m} = 1 \mid \mathcal{F}_{i,t-1}\right) = f_m\left(\mathbf{X}_{i,t-1}^{D}, \mathbf{X}_{i,t-1}^{E}\right),$$

where V_{it}^{m} denotes the breadth-, severity-, or PCA-based binary vulnerability state defined in Supplementary Appendix A, Section A.3. The domestic macro-fiscal information set is

$$\mathbf{X}_{i,t-1}^{D} = (D, B, R, E, G, \pi)'_{i,t-1},$$

and the external-sector information set is

$$\mathbf{X}_{i,t-1}^{E} = (CA, FX, \Delta REER, EX, IM)'_{i,t-1}.$$

All predictors are dated $t-1$. The exercise therefore estimates prospective one-year-ahead vulnerability probabilities rather than contemporaneous classifications. The mapping $f_m(\cdot)$ is predictive and is not assigned a structural or causal interpretation.

The raw developing-economy panel begins in 1980. Because several predictor variables require annual transformations and all forecasts use one-year lags, 1982 is the first admissible forecast year. The resulting forecasting sample is partitioned chronologically rather than randomly. Table B1.1 reports the principal temporal blocks used for development, calibration, evaluation, and post-evaluation sensitivity analysis.

Table B1.1. Temporal forecasting design and prepared-data contract

Analysis block	Period	Observations	Economies	Breadth events	Event share	Role
Development/model selection	1982–2005	2,219	123	748	0.337	Recursive specification selection
Internal calibration	2006–2009	493	124	83	0.168	Probability recalibration
Internal evaluation	2010–2017	989	125	204	0.206	Held-out temporal evaluation
External evaluation	2018–2023	719	124	197	0.274	Principal external evaluation
2024 sensitivity	2024	107	107	29	0.271	Post-evaluation sensitivity

Notes: Event counts refer to the baseline breadth definition using the fixed development-sample measurement parameters. Counts for the severity- and PCA-based outcomes differ because the outcome definitions differ. The development sample is used for specification selection; observations assigned to subsequent evaluation periods do not enter that selection exercise. The 2024 sample is excluded from the principal external evaluation.

B.1.1 Recursive development-period validation

Candidate-model selection is conducted within the 1982-2005 development period using expanding-window rolling-origin validation. Let $\mathcal{T}_f$ denote the historical training sample for fold f and $\mathcal{V}_f$ the subsequent validation block. The design satisfies

$$\max\{t: (i,t) \in \mathcal{T}_f\} < \min\{t: (i,t) \in \mathcal{V}_f\},$$

so that every observation used for estimation precedes every observation used for validation within a fold.

The three validation folds are reported in Table B1.2. The training window expands sequentially, while each validation block remains chronologically subsequent to its corresponding training sample.

Table B1.2. Expanding-window rolling-origin validation design

Fold	Training period	Validation period	Training observations	Validation observations
1	1982–1993	1994–1997	906	388
2	1982–1997	1998–2001	1,294	448
3	1982–2001	2002–2005	1,742	477

Notes: Observation counts refer to country-years entering the recursive development exercise. Outcome labels used for model selection are reconstructed separately within each fold. Consequently, fold-specific event counts need not coincide with event counts obtained by applying the fixed full-development-sample thresholds retrospectively to the same observations.

The recursive design applies not only to model estimation but also to construction of the dependent variable. This distinction is important because the vulnerability outcomes themselves depend on estimated quantiles, scale parameters, and, for the PCA definition, statistical-component parameters. Estimating those quantities once from the full 1982-2005 period before rolling validation would permit later development-period observations to influence labels assigned to earlier validation blocks.

For the breadth outcome, the fold-specific threshold for dimension k is therefore

$$c_{k,f} = \begin{cases} Q_{0.75}\left(Z_k \mid \mathcal{T}_f\right), & d_k = +1, \\ Q_{0.25}\left(Z_k \mid \mathcal{T}_f\right), & d_k = -1, \end{cases}$$

and the corresponding validation-period component indicator is

$$R_{kit}^{(f)} = \mathbf{1}\{d_k\left(Z_{kit} - c_{k,f}\right) \geq 0\}, \qquad (i,t) \in \mathcal{V}_f.$$

The fold-specific breadth score and outcome are

$$B_{it}^{(f)} = \sum_{k=1}^{8} R_{kit}^{(f)},$$

and

$$V_{it}^{B,(f)} = \mathbf{1}\left\{B_{it}^{(f)} \geq 3\right\}.$$

The same training-history restriction is applied to all parameters required by the severity and PCA constructions. For severity, the adverse thresholds, scale parameters, and severity cut-off are estimated from $\mathcal{T}_f$. For PCA, adversity-oriented location and scale parameters, the leading loading vector, sign normalization, and binary cut-off are estimated from $\mathcal{T}_f$. Each validation block is then transformed and labelled using those frozen training-fold parameters.

The recursive restriction also applies to predictor preprocessing. Clipping bounds, missing-value imputation parameters, and feature-scaling parameters, where required by the candidate algorithm, are estimated only from $\mathcal{T}_f$ and transferred unchanged to $\mathcal{V}_f$. The detailed preprocessing algorithms are reported in Section B.2.

B.1.2 Internal temporal evaluation

After model selection, the selected architecture is re-estimated using the complete 1982–2005 development sample. Its raw predicted probabilities are recalibrated using the distinct 2006–2009 calibration period, and the resulting forecasting pipeline is evaluated on the untouched 2010-2017 sample.

Denoting the three blocks by $\mathcal{D}$, $\mathcal{C}$, and $\mathcal{E}_I$,

$$\mathcal{D} = \{1982, \dots, 2005\},$$

$$\mathcal{C} = \{2006, \dots, 2009\},$$

and

$$\mathcal{E}_I = \{2010, \dots, 2017\},$$

the temporal ordering satisfies

$$\max(\mathcal{D}) < \min(\mathcal{C}) \leq \max(\mathcal{C}) < \min(\mathcal{E}_I).$$

Neither candidate-model selection nor model fitting uses observations from $\mathcal{C}$ or $\mathcal{E}_I$, and the internal evaluation observations are not used for probability recalibration.

For the locked internal evaluation, the vulnerability-measurement parameters are estimated from the complete 1982-2005 development sample and then held fixed. Fold-specific reconstruction is required during recursive model selection because different validation blocks lie inside the development period; once specification selection is complete, the full development-sample measurement parameters define the outcomes used in subsequent calibration and evaluation.

B.1.3 External temporal evaluation

The external evaluation is designed to assess temporal generalization after additional historical information becomes available without reopening model selection. The model family and hyperparameters selected in the development exercise remain fixed. The locked architecture is re-estimated using information available through 2013, probability recalibration is performed using 2014-2017 observations, and final performance is evaluated over 2018–2023.

Let

$$\mathcal{D}_E = \{1982, \ldots, 2013\},$$

$$\mathcal{C}_E = \{2014, \ldots, 2017\},$$

and

$$\mathcal{E}_E = \{2018, \ldots, 2023\}.$$

The external design therefore satisfies

$$\max(\mathcal{D}_E) < \min(\mathcal{C}_E) \leq \max(\mathcal{C}_E) < \min(\mathcal{E}_E).$$

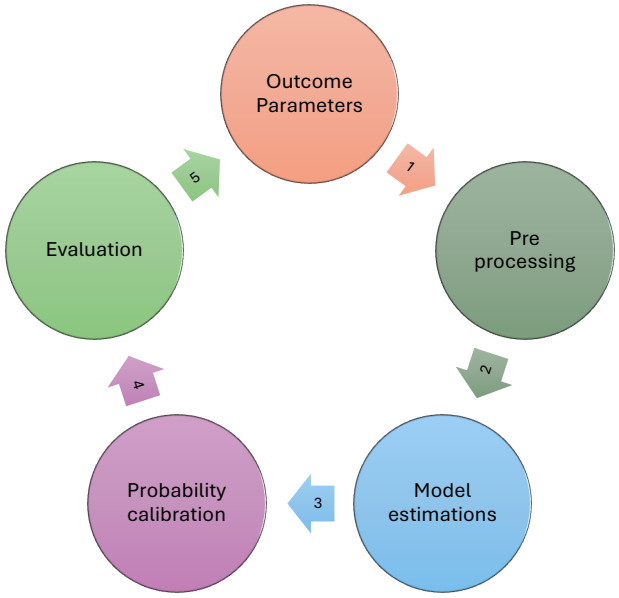
The additional 2006-2013 observations enter external model re-estimation because they would have been historically available before the external forecast period. They do not trigger a new model search. Similarly, the 2014-2017 observations are used for recalibration only and are not used to evaluate the final external forecasts.

The measurement parameters defining the breadth, severity, and PCA outcomes remain fixed at their original 1982-2005 development-sample estimates when constructing outcomes for the later calibration and evaluation periods. This maintains a common measurement rule over time and prevents the definition of vulnerability from adapting retrospectively to conditions observed in the evaluation sample.

B.1.4 Leakage-control protocol

The forecasting design separates five operations that can otherwise generate look-ahead bias: outcome construction, predictor preprocessing, candidate-model selection, model estimation, and probability calibration. For an evaluation observation in year t, no observation dated later than the admissible historical training or calibration window is permitted to affect any of these operations.

The resulting information restrictions can be summarized as:



with every fitted object estimated from a chronologically prior sample appropriate to its role. Random country-year train-test partitions are not used. Country observations may appear in both earlier estimation periods and later evaluation periods, as required by a pooled panel forecasting design, but future observations for the same economy cannot enter the estimation of forecasts assigned to earlier years.

The 2024 observations are excluded from model selection and from the principal internal and external evaluations. They are retained only for the separate post-evaluation sensitivity exercise reported in Section B.10.

B.2 Fold-specific preprocessing, candidate-model search, and specification selection

All predictor preprocessing is nested within the chronological training sample associated with each rolling-origin fold. No transformation parameter is estimated from validation observations. For fold f, let $\mathcal{T}_f$ and $\mathcal{V}_f$ denote the corresponding training and validation samples defined in Section

B.1. A preprocessing operator $\mathcal{P}_f(\cdot)$ is estimated exclusively from $\mathcal{T}_f$ and subsequently applied unchanged to both the training and validation predictor matrices.

For each continuous predictor x_j, the training-fold clipping limits are

$$\ell_{jf} = Q_{0.01}(x_j \mid \mathcal{T}_f), \qquad u_{jf} = Q_{0.99}(x_j \mid \mathcal{T}_f),$$

and the transformed value is

$$x_{ijt}^{c} = \min\{u_{jf}, \max(\ell_{jf}, x_{ijt})\}.$$

Thus, extreme values are clipped rather than deleted, preserving country-year observations while limiting the influence of unusually large realizations on model estimation.

Missing predictor values are subsequently replaced with the corresponding training-fold median,

$$x_{ijt}^{imp} = \begin{cases} x_{ijt}^{c}, & x_{ijt}^{c} \text{ observed,} \\ \tilde{x}_{jf}, & x_{ijt}^{c} \text{ missing,} \end{cases}$$

where

$$\tilde{x}_{jf} = \text{median}(x_j^{c} \mid \mathcal{T}_f).$$

No missingness indicators are added. The same fitted medians are applied to the corresponding validation observations.

Algorithms whose estimation depends on predictor scale additionally receive standardization based exclusively on the training fold,

$$x_{ijt}^{std} = \frac{x_{ijt}^{imp} - \mu_{jf}}{\sigma_{jf}},$$

where μ_{jf} and σ_{jf} are estimated after clipping and imputation within $\mathcal{T}_f$. Standardization is applied to logistic regression, elastic-net logistic regression, support-vector machines, k-nearest neighbours, Gaussian naive Bayes, and the feedforward neural-network specifications. Tree-based and boosting estimators receive fold-specific clipping and median imputation but are not standardized.

This transformation sequence can be written compactly as

$$\mathcal{P}_f = \mathcal{S}_f \circ \mathcal{I}_f \circ \mathcal{C}_f,$$

where $\mathcal{C}_f$ denotes 1st -99th percentile clipping, $\mathcal{I}_f$ denotes median imputation, and $\mathcal{S}_f$ denotes standardization when required by the candidate estimator. All three operators are fitted using $\mathcal{T}_f$ only.

B.2.1 Candidate-model library

The model library contains 20 prespecified configurations from 13 supervised-learning families. The objective of the search is comparative predictive specification selection rather than the introduction of a new forecasting algorithm. The same eleven lagged predictors are supplied to every candidate, and all candidates are evaluated on the same three rolling-origin folds described in Table B1.2.

Table B2.1 reports the complete candidate library and the principal hyperparameters fixed for the final model-selection exercise. The random seed is 42 throughout. In the final run, random forests and extremely randomized trees use 600 trees, boosting specifications use 450 boosting iterations or estimators, and neural networks permit a maximum of 1,800 optimization iterations.

Table B2.1. Complete candidate-model library and fixed hyperparameters

Candidate	Model family	Principal specification	Standardized inputs
Logit, $C = 0.2$	Logistic regression	$C = 0.2$; balanced class weights; max. 4,000 iterations	Yes
Logit, $C = 1.0$	Logistic regression	$C = 1.0$; balanced class weights; max. 4,000 iterations	Yes
Elastic net, $C = 0.2$	Elastic-net logistic regression	$C = 0.2$; l_1 ratio $= 0.5$; saga; balanced class weights	Yes
Elastic net, $C = 1.0$	Elastic-net logistic regression	$C = 1.0$; l_1 ratio $= 0.5$; saga; balanced class weights	Yes
SVM, $C = 0.5$	Support-vector machine	RBF kernel; $C = 0.5$; probability output; balanced class weights	Yes
SVM,$C = 1.0$	Support-vector machine	RBF kernel; $C = 1.0$; probability output; balanced class weights	Yes
KNN, $k = 15$	k-nearest neighbours	15 neighbours	Yes
Gaussian naive Bayes	Naive Bayes	Gaussian likelihood	Yes
MLP, 16	Artificial neural network	One hidden layer, 16 ReLU units; $\alpha = 0.01$; batch 64; learning rate 0.001	Yes
MLP, 32–16	Artificial neural network	Hidden layers 32 and 16; ReLU; $\alpha = 0.01$; batch 64; learning rate 0.001	Yes
Decision tree, depth 5	Decision tree	Maximum depth 5; minimum leaf size 20; balanced class weights	No

Candidate	Model family	Principal specification	Standardized inputs
Random forest, depth 4	Random forest	600 trees; depth 4; minimum leaf size 10; $\sqrt{p}$features; balanced class weights	No
Random forest, depth 7	Random forest	600 trees; depth 7; minimum leaf size 10; $\sqrt{p}$features; balanced class weights	No
Extra Trees, depth 6	Extremely randomized trees	600 trees; depth 6; minimum leaf size 10;$\sqrt{p}$features; balanced class weights	No
Gradient boosting, depth 2	Gradient boosting	450 estimators; learning rate 0.03; maximum depth 2	No
Gradient boosting, depth 3	Gradient boosting	450 estimators; learning rate 0.03; maximum depth 3	No
Histogram gradient boosting	Histogram gradient boosting	450 iterations; learning rate 0.04; maximum 15 leaves; $L_2 = 1$	No
AdaBoost	AdaBoost	450 estimators; learning rate 0.03	No
XGBoost, depth 2	XGBoost	450 trees; depth 2; learning rate 0.03; subsample 0.8; column subsample 0.8; minimum child weight 5; $L_2 = 1$	No
XGBoost, depth 3	XGBoost	450 trees; depth 3; learning rate 0.03; subsample 0.8; column subsample 0.8; minimum child weight 5; $L_2 = 1$	No

Notes: Every specification uses the same eleven one-year-lagged predictors. All predictors are clipped at training-fold 1st and 99th percentiles and median-imputed before estimation. Standardization is applied only where indicated. Neural-network early stopping is disabled so that no randomly selected internal validation set is introduced within a historical training fold. The candidate configurations are fixed before evaluation of the 2010–2017 and 2018–2023 samples.

The neural-network candidates are deliberately compact because the forecasting problem is low-dimensional tabular classification rather than a high-dimensional image or sequence problem. Their hidden layers use rectified linear unit activation, the Adam optimizer, L_2 regularization with $\alpha = 0.01$, and logistic probability output. Random early stopping is disabled because such a procedure would introduce an additional internal split that does not respect the imposed chronological validation design.

B.2.2 Model-selection criterion

Each candidate $r \in \mathcal{M}$ is evaluated separately in the three expanding-window folds. Let

$$AUC_{rf}, \qquad AP_{rf}, \qquad BS_{rf}$$

denote the ROC AUC, average precision, and Brier score of candidate \(r\) in validation fold \(f\).

The fold averages are

$$\overline{AUC}_r = \frac{1}{F}\sum_{f=1}^{F} AUC_{rf},$$

$$[\overline{AP}_r = \frac{1}{F}\sum_{f=1}^{F} AP_{rf},$$

and

$$\overline{BS}_r = \frac{1}{F}\sum_{f=1}^{F} BS_{rf},$$

with $F = 3$.

Specification selection follows the prespecified lexicographic rule

$$\hat{r} = \arg\operatorname{lexmax}_{r\in\mathcal{M}} \left\{\overline{AUC}_r, \overline{AP}_r, \min_f AUC_{rf}, -\overline{BS}_r\right\}.$$

Mean ROC AUC is therefore the primary selection statistic. Mean average precision is the first tie-breaker, followed by the worst fold-specific ROC AUC and, finally, lower mean Brier loss. The rule is applied using development-period validation results only. No calibration-period, internal-evaluation, external-evaluation, or 2024 observations enter model selection.

Table B2.2 reports the selected specification together with the principal comparison models presented in the main manuscript. The depth-7 random forest records the highest mean rolling-origin ROC AUC, 0.876, and is consequently selected. Its advantage over depth-2 XGBoost is only 0.001, while Extra Trees and Gradient Boosting both attain mean AUCs of approximately 0.870. XGBoost records marginally higher average precision and a lower mean Brier score than the selected random forest. The validation evidence therefore supports a competitive group of nonlinear ensemble models rather than dominance by a single algorithm.

Table B2.2. Rolling-origin performance of the selected and principal comparison specifications

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

Specification	Mean AUC	Worst-fold AUC	Mean AP	Mean Brier
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 the three expanding-window rolling-origin validation folds within 1982–2005. Outcome parameters are reconstructed within each fold using the corresponding historical training sample. Mean ROC AUC is the primary selection criterion, followed sequentially by mean average precision, minimum fold-specific ROC AUC, and mean Brier loss. The full candidate library is reported in Table B2.1.

Figure B2.1 provides the corresponding model-family comparison. The dispersion across historical folds is non-negligible for several candidates, reinforcing the use of repeated temporal validation rather than selection from a single holdout period. The figure also shows that several nonlinear ensemble families occupy a relatively narrow upper range of development-period discrimination.

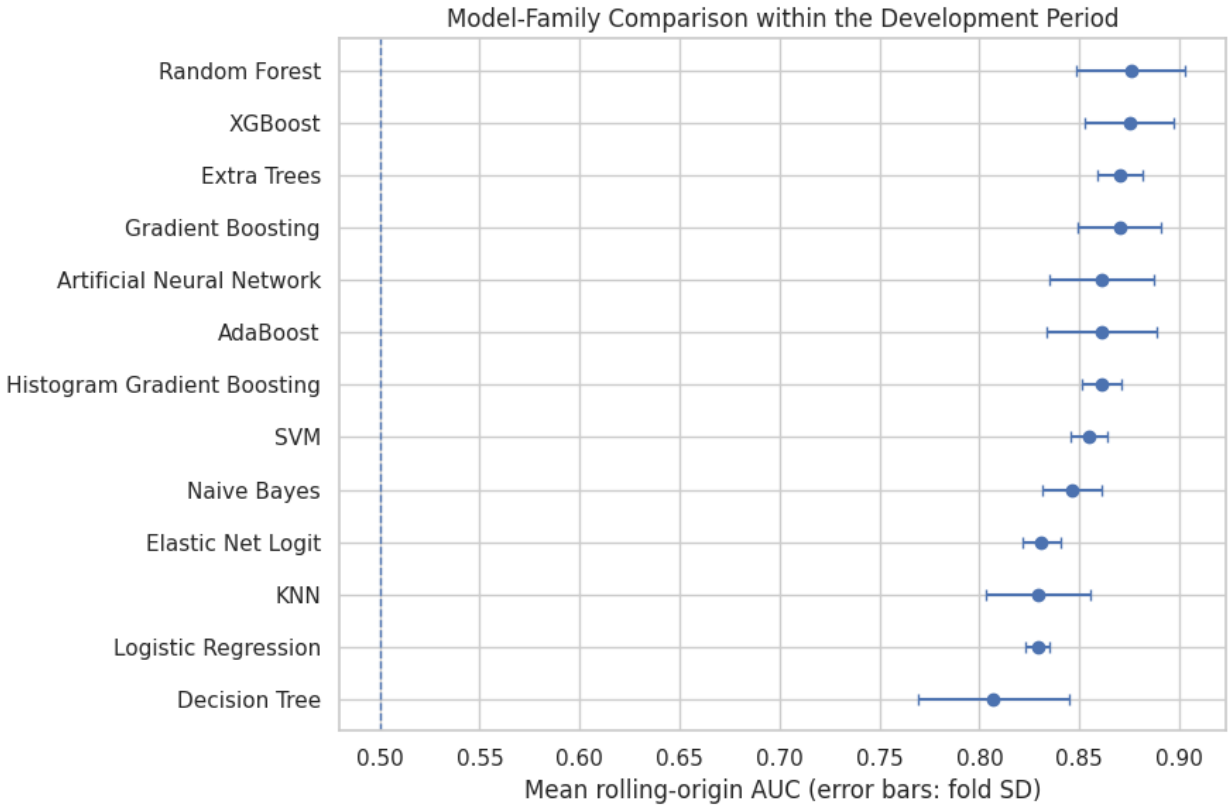


Figure B2.1. Rolling-origin model-family comparison within the development period

Notes: *Points report mean ROC AUC for the highest-ranked specification within each candidate-model family; horizontal error bars report the standard deviation of ROC AUC across the three rolling-origin folds. The vertical reference corresponds to ROC AUC* $= 0.5$. *Model selection itself uses the complete candidate set and the lexicographic criterion defined above.*

Once selected, the random-forest architecture is locked. The model family and hyperparameters are not re-selected using the 2010-2017 internal evaluation, the 2018-2023 external evaluation, or alternative vulnerability definitions. For robustness analysis, the same depth-7 random-forest architecture is re-estimated under the breadth, severity, PCA, and supplementary outcome

constructions so that differences in forecast performance reflect changes in the dependent-variable definition rather than renewed algorithm selection.

B.3 Probability recalibration and forecast-performance metrics

The selected forecasting specification produces raw class probabilities that need not be well calibrated in absolute probability terms. Probability recalibration is therefore separated chronologically from model estimation and final evaluation. For the internal exercise, the random-forest specification is estimated on 1982-2005, recalibrated using 2006-2009, and evaluated on 2010-2017. For the external exercise, the same locked specification is re-estimated through 2013, recalibrated using 2014-2017, and evaluated on 2018-2023. In neither case are evaluation-period outcomes used to estimate the calibration transformation.

B.3.1 Logistic probability recalibration

Let $\hat{p}_{it}$ denote the raw probability generated by the fitted classifier. Recalibration is implemented by fitting a logistic model to the log odds of the raw probability in the designated calibration sample:

$$\text{logit}(\tilde{p}_{it}) = \alpha_c + \beta_c \text{logit}(\hat{p}_{it}),$$

where

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

For numerical stability, raw probabilities are restricted to the interval $[10^{-6}, 1-10^{-6}]$ before the logit transformation is applied. The fitted calibration intercept $\hat{\alpha}_c$ and slope $\hat{\beta}_c$ are then transferred unchanged to the subsequent evaluation period.

The calibrated probability is therefore

$$\tilde{p}_{it} = \frac{1}{1 + \exp\left[-\left(\hat{\alpha}_c + \hat{\beta}_c \,\text{logit}(\hat{p}_{it})\right)\right]}.$$

The calibration intercept primarily adjusts the overall location of the probability distribution, while the slope adjusts its dispersion. Because the transformation is monotonic when the fitted slope is positive, observation rankings are preserved. ROC AUC and average precision therefore remain

unchanged by recalibration, whereas the Brier score, log loss, expected calibration error, and classifications at a fixed probability threshold can change.

Recalibration improves probability accuracy and calibration in both evaluation periods. In the internal sample, the Brier score declines from 0.127 to 0.120, while expected calibration error decreases from 0.088 to 0.039. In the external sample, the corresponding reductions are from 0.150 to 0.141 and from 0.102 to 0.057, respectively. As expected, AUC and average precision remain unchanged.

Table B3.1. Raw and recalibrated predictive performance

Evaluation period	Probability	AUC	AP	Brier	Log loss	ECE_{10}	Precision	Recall	F1	Balanced accuracy	False-alarm rate	Missed-vulnerability rate
Internal, 2010–2017	Raw	0.838	0.595	0.127	0.409	0.088	0.430	0.799	0.559	0.762	0.275	0.201
Internal, 2010–2017	Logistic recalibrated	0.838	0.595	0.120	0.380	0.039	0.500	0.618	0.553	0.729	0.161	0.382
External, 2018–2023	Raw	0.819	0.682	0.150	0.466	0.102	0.441	0.822	0.574	0.715	0.393	0.178
External, 2018–2023	Logistic recalibrated	0.819	0.682	0.141	0.444	0.057	0.538	0.675	0.599	0.728	0.218	0.325

Notes: *Classification metrics use the 0.30 operating threshold. Recalibration is estimated only from the designated calibration period and then applied unchanged to the subsequent evaluation observations. AP denotes average precision. ECE_{10} denotes expected calibration error based on ten equal-width probability bins. Detailed sensitivity to the operating threshold is reported in Section B.8.*

The movement in threshold-dependent statistics in Table B3.1 is not evidence of a change in discrimination. Rather, recalibration changes the location of probability forecasts relative to the fixed 0.30 operating threshold. For example, the external false-alarm rate falls from 0.393 to 0.218 following recalibration, while recall falls from 0.822 to 0.675. This illustrates why discrimination, calibration, and threshold-dependent classification performance are evaluated separately.

B.3.2 Forecast-performance metrics

Forecast performance is assessed along three distinct dimensions: discrimination, probability accuracy, and calibration. ROC AUC evaluates ranking ability and is invariant to event prevalence and monotonic transformations of the forecast score. Average precision is used as the precision-recall summary statistic because vulnerability is not equally prevalent across outcome definitions and evaluation periods. Its no-skill reference equals the event share in the relevant evaluation sample.

Average precision is calculated as

$$\mathrm{AP} = \sum_{n} (R_n - R_{n-1})\, P_n,$$

where P_n and R_n denote precision and recall at successive probability thresholds.

Probability accuracy is summarized primarily by the Brier score,

$$\mathrm{BS} = \frac{1}{N} \sum_{j=1}^{N} \left(V_j - \tilde{p}_j\right)^2,$$

where lower values indicate more accurate probabilistic forecasts.

Logarithmic loss is retained as an additional proper scoring rule:

$$\mathrm{LL} = -\frac{1}{N} \sum_{j=1}^{N} \left[V_j \log\left(\tilde{p}_j\right) + \left(1 - V_j\right) \log\left(1 - \tilde{p}_j\right)\right].$$

Calibration is summarized by expected calibration error using ten equal-width probability bins. Let B_b denote probability bin b, n_b its number of observations, $\overline{p}_b$ its mean predicted probability, and $\overline{V}_b$ its observed vulnerability frequency. Then

$$ECE_{10} = \sum_{b=1}^{10} \frac{n_b}{N} \left|\overline{p}_b - \overline{V}_b\right|.$$

Because ECE_{10} depends on sample size and binning choices, it is treated as a descriptive calibration diagnostic rather than as an inferential statistic.

At the prespecified 0.30 operating threshold, the appendix also reports precision, recall, F1, balanced accuracy, the false-alarm rate, and the missed-vulnerability rate. The false-alarm rate is $FP/(FP + TN)$, while the missed-vulnerability rate is $FN/(FN + TP) = 1 - \mathrm{recall}$. These operating-point measures are secondary to the threshold-free discrimination and probability metrics because the appropriate surveillance threshold ultimately depends on the relative institutional costs of missed vulnerabilities and false warnings.

Figure B3.1 presents the ROC and precision-recall curves for the calibrated baseline forecasts. The internal ROC AUC is 0.838 and the external ROC AUC is 0.819. Average precision is 0.595

internally and rises to 0.682 externally, against an external event prevalence of approximately 0.274. The model therefore retains substantial ranking information in the later external period despite the stricter temporal separation from model selection.

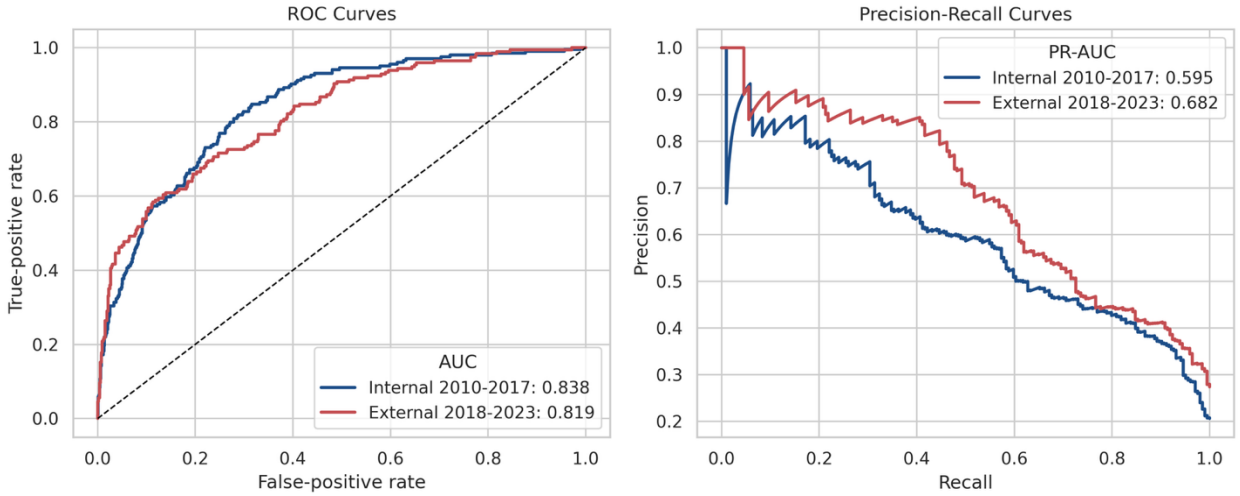


Figure B3.1. Internal and external discrimination curves

Notes: *Curves are based on recalibrated probabilities from the selected depth-7 random forest. The ROC panel reports ROC AUC. The precision-recall panel reports the average-precision statistic; the figure legend uses "PR-AUC" as the plotting label for this quantity.*

B.3.3 Calibration diagnostics

Calibration intercept and slope provide a complementary diagnostic of the probability scale. They are estimated retrospectively from the evaluation-period predictions solely for diagnostic purposes and are not used to modify those forecasts. A perfectly calibrated set of predictions would have an intercept of zero and a slope of one.

Table B3.2. Calibration intercept and slope

Evaluation period	Probability	Calibration intercept	Calibration slope
Internal, 2010–2017	Raw	-0.431	1.308
Internal, 2010–2017	Logistic recalibrated	-0.185	0.905
External, 2018–2023	Raw	-0.531	1.152
External, 2018–2023	Logistic recalibrated	-0.181	0.817

Notes: Calibration intercepts and slopes are estimated retrospectively on the evaluation-period predictions as diagnostics only. They do not enter the probability-recalibration procedure used to generate those predictions.

As shown in Table B3.2, recalibration moves both evaluation samples closer to these reference values. For the internal evaluation, the calibration intercept changes from -0.431 to -0.185 and the slope from 1.308 to 0.905. For the external evaluation, the corresponding values change from

-0.531 to -0.181 and from 1.152 to 0.817. The remaining deviations indicate that chronological recalibration improves, but does not eliminate, probability miscalibration under temporal distributional change.

Figure B3.2 compares mean calibrated probabilities with observed vulnerability incidence across probability deciles. Unlike ECE_{10}, which uses ten fixed-width bins over $[0,1]$, the graphical calibration diagnostic divides observations into approximately equal-sized groups according to predicted probability. This provides greater resolution in probability ranges containing most observations.

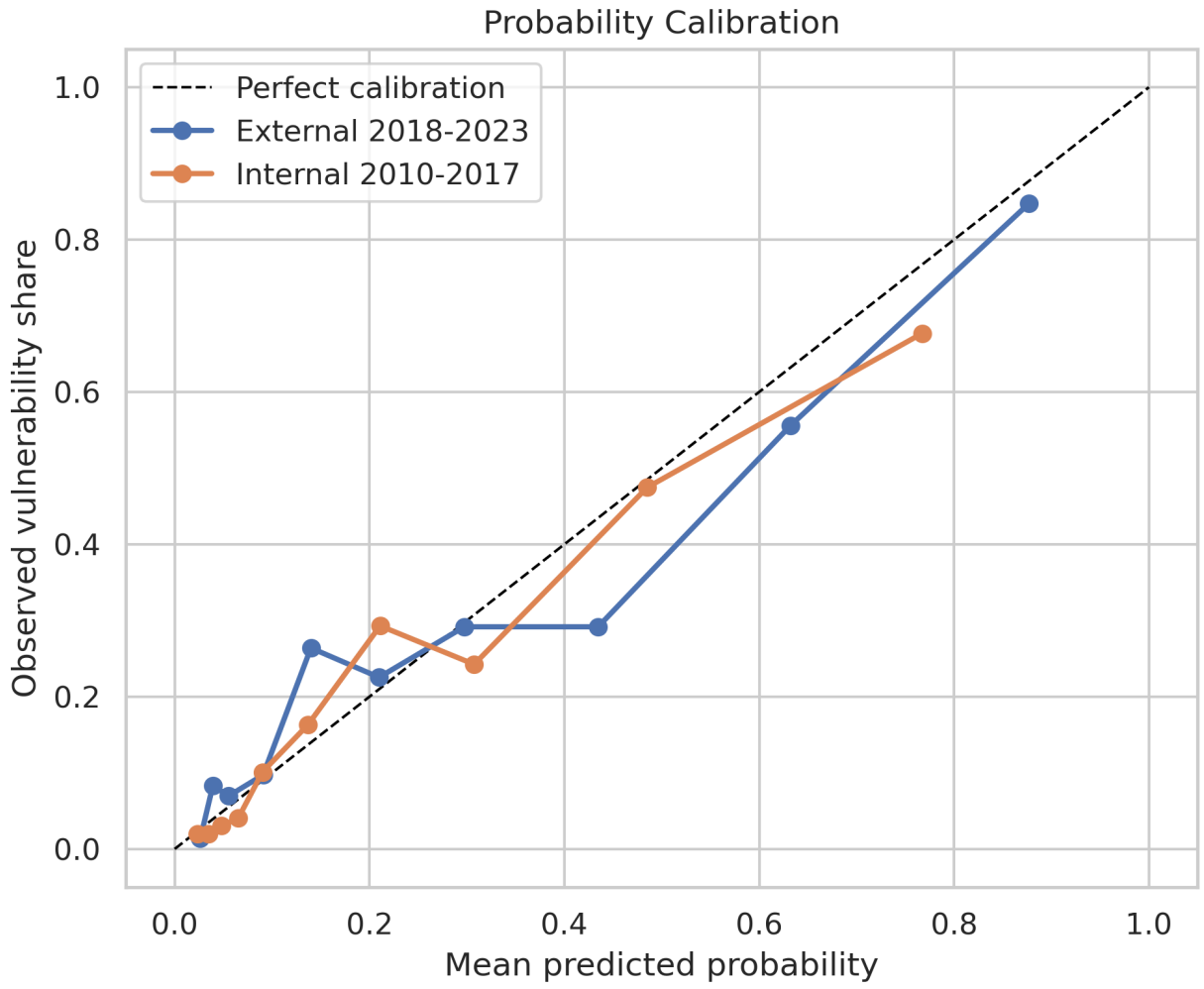


Figure B3.2. Internal and external probability calibration

Notes: *Points report mean predicted probability and observed vulnerability incidence within deciles of the calibrated probability distribution. The 45-degree line represents perfect calibration. Decile groups are constructed separately within each evaluation period.*

Table B3.3 reports the underlying decile-level values. Calibration is generally close in the lower and upper probability ranges, although some intermediate bins show larger discrepancies. These deviations are consistent with the positive aggregate ECE_{10} values and with the calibration slopes below one after recalibration.

Table B3.3. Decile-level calibration diagnostics

Evaluation period	Mean predicted probability	Observed vulnerability share	Observations
Internal, 2010–2017	0.023	0.020	99
Internal, 2010–2017	0.035	0.020	99
Internal, 2010–2017	0.048	0.030	99
Internal, 2010–2017	0.065	0.040	99
Internal, 2010–2017	0.090	0.101	99
Internal, 2010–2017	0.137	0.163	98
Internal, 2010–2017	0.211	0.293	99
Internal, 2010–2017	0.307	0.242	99
Internal, 2010–2017	0.485	0.475	99
Internal, 2010–2017	0.767	0.677	99
External, 2018–2023	0.026	0.014	72
External, 2018–2023	0.039	0.083	72
External, 2018–2023	0.055	0.069	72
External, 2018–2023	0.091	0.097	72
External, 2018–2023	0.140	0.264	72
External, 2018–2023	0.210	0.225	71
External, 2018–2023	0.297	0.292	72
External, 2018–2023	0.435	0.292	72
External, 2018–2023	0.632	0.556	72
External, 2018–2023	0.877	0.847	72

Notes: Observations are divided into ten approximately equal-sized groups using quantiles of the recalibrated predicted-probability distribution separately within each evaluation period. These decile groups are used for the graphical calibration diagnostic in Figure B3.2 and differ from the fixed-width probability bins used to calculate ECE_{10}.

B.4 Persistence and Conventional Forecasting Benchmarks

Fiscal vulnerability is inherently persistent because public debt, fiscal balances, revenue capacity, and external-sector conditions often exhibit considerable serial dependence. Consequently, strong out-of-sample forecast performance does not necessarily imply that a multivariate forecasting model provides substantial information beyond the vulnerability state already observed in the preceding year. A model may discriminate effectively between vulnerable and non-vulnerable country-years primarily because adverse conditions persist over time.

To distinguish overall forecastability from incremental predictive content, the selected random forest is evaluated against four benchmarks: fitted lagged-score persistence, conventional multivariate logistic regression, lagged binary-state persistence, and a historical-event-rate forecast. Particular emphasis is placed on fitted lagged-score persistence because it directly tests whether the broader macro-fiscal information set contributes predictive information beyond the previous value of the vulnerability measure.

All benchmarks follow the chronological estimation and evaluation design described in Section B.1. The internal evaluation uses observations from 2010–2017, while the external evaluation uses 2018–2023. Model parameters are estimated using the corresponding historical estimation samples, and probability recalibration is performed using the designated calibration periods where applicable. Evaluation-period observations are excluded from model estimation and probability recalibration.

B.4.1 Benchmark specifications

Fitted lagged-score persistence

Let $S_{it}^{(m)}$ denote the continuous vulnerability score underlying binary vulnerability definition m, where m indexes breadth, severity, or the PCA-based statistical component. Specifically, $S_{it}^{(m)}$ corresponds to the breadth count, the threshold-distance severity index, or the first principal-component score, respectively.

The fitted persistence benchmark estimates the probability of vulnerability using only the preceding year's continuous vulnerability score:

$$\Pr\left(Y_{it}^{(m)} = 1 \mid S_{i,t-1}^{(m)}\right) = \Lambda\left(\alpha^{(m)} + \beta^{(m)} S_{i,t-1}^{(m)}\right),$$

where $\Lambda(z) = 1/(1 + e^{-z})$ denotes the logistic function, and $\alpha^{(m)}$ and $\beta^{(m)}$ are estimated using historical observations available before the relevant evaluation period.

Unlike a benchmark based solely on the preceding binary vulnerability state, this specification retains information about the intensity of vulnerability. For example, under the breadth construction, two country-years with scores of three and six are both classified as vulnerable, but the continuous score distinguishes their degrees of simultaneous macro-fiscal stress.

The lagged score is retained only when an observation for the same economy is available in the immediately preceding calendar year. The fitted persistence probabilities are recalibrated using the corresponding historical calibration period, consistent with the forecasting protocol applied to the selected multivariate model.

The persistence benchmark provides a demanding reference because it uses a compact historical summary of the same underlying macro-fiscal conditions from which the vulnerability outcome is constructed. The comparison therefore evaluates incremental forecast performance beyond this summary rather than the complete absence of predictive information in the previous year's macroeconomic variables.

Conventional multivariate logistic regression

The conventional econometric benchmark uses the same eleven lagged predictors supplied to the selected random forest but imposes a linear specification in the log odds of vulnerability:

$$\Pr\left(Y_{it}^{(m)} = 1 \mid X_{i,t-1}\right) = \Lambda\left(\gamma_0^{(m)} + X_{i,t-1}{}'\boldsymbol{\gamma}^{(m)}\right).$$

The predictor vector includes the fiscal, domestic macroeconomic, and external-sector variables defined in Section B.1. Predictor clipping, median imputation, and standardization are estimated exclusively from the corresponding historical training observations. The logistic specification uses the development-selected regularization configuration with $C = 0.2$.

This benchmark evaluates whether the flexible nonlinear specification improves predictive performance relative to a conventional model using the same information set.

Lagged binary-state persistence

A simpler persistence benchmark forecasts the current vulnerability state directly from the previous binary classification:

$$\hat{p}_{it,\text{binary}}^{(m)} = Y_{i,t-1}^{(m)}.$$

This benchmark assigns a probability of one when the economy was vulnerable in the preceding year and zero otherwise. It therefore assumes perfect continuation of the previous state and does not estimate transition probabilities.

Although restrictive, binary-state persistence provides a useful reference for assessing how much discrimination arises from the stability of the vulnerability classification itself.

Historical-event-rate benchmark

The final benchmark assigns every evaluation observation the historical vulnerability frequency observed in the corresponding estimation sample:

$$\hat{p}_{it,\text{historical}}^{(m)} = \frac{1}{N_{\mathcal{T}}} \sum_{(j,s)\in\mathcal{T}} Y_{js}^{(m)},$$

where $\mathcal{T}$ denotes the relevant historical estimation sample and $N_{\mathcal{T}}$ is its number of observations.

Because this benchmark generates a constant probability within an evaluation exercise, it contains no cross-country or intertemporal ranking information and has an ROC AUC of 0.5 whenever both outcome classes are observed. It serves as a reference for assessing probability forecasts against unconditional historical incidence.

B.4.2 Comparative forecasting performance

Table B4.1 compares the selected multivariate model with the four benchmarks in the internal and external evaluation periods. Performance is summarized using ROC AUC, average precision (AP), and the Brier score. The persistence benchmarks have slightly smaller evaluation samples because their construction requires a valid adjacent-year lag of the vulnerability score or state.

The internal evaluation indicates that fitted persistence provides a substantially more demanding benchmark than the unconditional historical-event-rate forecast. The fitted persistence model attains an AUC of 0.823, compared with 0.838 for the selected random forest. Although the random forest has superior discrimination, the difference is relatively modest, indicating that the previous breadth score already contains considerable information about vulnerability in the subsequent year. The random forest demonstrates a clearer advantage in average precision, recording 0.595 compared with 0.516 for fitted persistence. Its Brier score of 0.120 is also slightly below the persistence benchmark's 0.121. The relatively small difference in probability loss indicates that incremental gains depend on the particular evaluation criterion and should not be inferred from discrimination alone.

Table B4.1. Baseline breadth vulnerability: selected model versus conventional and persistence benchmarks

Evaluation period	Forecasting model	Obs.	AUC	AP	Brier
Internal 2010–2017	Selected random forest	989	0.838	0.595	0.120
	Logistic regression	989	0.778	0.526	0.133
	Fitted lagged-score persistence	987	0.823	0.516	0.121
	Binary-state persistence	987	0.742	0.425	0.175
	Historical event rate	989	0.500	0.206	0.181
External 2018–2023	Selected random forest	719	0.819	0.682	0.141
	Logistic regression	719	0.799	0.643	0.149
	Fitted lagged-score persistence	718	0.797	0.603	0.150
	Binary-state persistence	718	0.713	0.458	0.224
	Historical event rate	719	0.500	0.274	0.199

Notes: The baseline outcome defines vulnerability as the activation of at least three of eight adverse macro-fiscal dimensions. AUC denotes the area under the receiver operating characteristic curve; AP denotes average precision. Lower Brier scores indicate greater probability accuracy. Fitted lagged-score persistence and binary-state persistence require valid adjacent-year score or state observations. Comparisons of incremental forecast performance are conducted on matched observations, as documented in Section B.5. Reported statistics are rounded to three decimal places.

The external evaluation provides a more demanding assessment of temporal generalization. The selected model achieves an AUC of 0.819 and AP of 0.682, compared with 0.797 and 0.603 for fitted persistence. Probability accuracy also favours the multivariate specification: the selected model records a Brier score of approximately 0.141 compared with 0.150 for fitted persistence.

When the comparison is restricted to the 718 country-years with forecasts from both models, the selected specification achieves a Brier score of 0.14144 against 0.15002 for fitted persistence. The corresponding Brier skill score is

$$\mathrm{BSS} = 1 - \frac{\mathrm{Brier}_{\mathrm{selected}}}{\mathrm{Brier}_{\mathrm{persistence}}} = 0.0572.$$

Thus, the selected multivariate model reduces out-of-sample quadratic probability loss by approximately 5.7% relative to fitted persistence under the baseline breadth definition.

Conventional logistic regression also remains competitive in the external period, recording an AUC of 0.799 and AP of 0.643. The random forest therefore provides an improvement over the linear specification, although the magnitude of the gain is smaller than would be suggested by a comparison against an unconditional benchmark.

Binary-state persistence is considerably less competitive, with external AUC of 0.713 and AP of 0.458. This contrast is informative because it demonstrates that the intensity of prior vulnerability, represented by the continuous score, carries considerably more information than the binary classification alone. Consequently, comparisons relying exclusively on lagged binary states could exaggerate the incremental forecasting advantage of multivariate models.

These comparisons establish two distinct findings. First, vulnerability exhibits substantial persistence, and a model based only on the preceding continuous score achieves strong discrimination. Second, the broader macro-fiscal forecasting system retains additional predictive content beyond that persistence, particularly in average precision and external-period probability accuracy.

However, the magnitude and statistical reliability of these incremental gains cannot be established from separate performance statistics alone. Because the forecast errors and observed outcomes are correlated across repeated observations within economies, formal inference requires paired comparisons on identical country-year samples with country-clustered uncertainty estimation. These comparisons are developed in Section B.5.

B.5 Forecast Uncertainty and Paired Model Comparisons

Strong out-of-sample forecast performance does not, by itself, establish that differences between competing models are statistically distinguishable from sampling variation. This issue is particularly important in the present application because the evaluation data comprise repeated annual observations for the same economies. Forecast errors may be correlated within countries owing to persistent macroeconomic conditions, fiscal-policy regimes, and other country-specific characteristics. Treating individual country-year observations as independent could consequently understate uncertainty.

Sampling uncertainty is therefore assessed using 1,000 country-clustered bootstrap replications. The analysis distinguishes uncertainty surrounding the absolute performance of the selected forecasting specification from uncertainty surrounding its incremental performance relative to competing benchmarks. For the latter, model comparisons are conducted on identical country-year observations, ensuring that differences in predictive performance are not driven by changes in sample composition.

The uncertainty estimates are conditional on the fitted forecasting pipelines. They quantify sampling variation within the held-out evaluation periods but do not incorporate uncertainty from renewed outcome construction, model selection, or parameter re-estimation within each bootstrap replication.

B.5.1 Country-clustered bootstrap uncertainty

Let $i = 1, \dots, G$index economies represented in a given evaluation sample, and let $\mathcal{E}_i$ denote the set of available evaluation-period observations for economy \(i\). The country-clustered bootstrap draws G economies with replacement from the observed set and retains all evaluation observations associated with each selected economy.

For bootstrap replication $b = 1, \dots, 1000$, denote the resampled evaluation data by

$$\mathcal{E}^{*(b)} = \bigcup_{g=1}^{G} \mathcal{E}_{i_g^{*(b)}},$$

where $i_g^{*(b)}$ denotes the economy selected at bootstrap position g. Economies selected multiple times contribute their observations with the corresponding multiplicity.

This procedure preserves the within-country dependence structure of the observed annual panel while allowing the composition of economies represented in each bootstrap replication to vary. It also accommodates the unbalanced nature of the analytical panel, since economies contribute different numbers of available observations.

For a forecast-performance statistic θ, such as ROC AUC, average precision, or the Brier score, the statistic is recomputed in each bootstrap sample:

$$\hat{\theta}^{*(b)} = T\left(\mathcal{E}^{*(b)}\right).$$

The 95% confidence interval is obtained from the empirical 2.5th and 97.5th percentiles of the bootstrap distribution:

$$CI_{0.95}(\hat{\theta}) = \left[Q_{0.025}(\hat{\theta}^*), Q_{0.975}(\hat{\theta}^*)\right].$$

Here, Q_τ denotes the empirical quantile at probability τ. Point estimates are calculated from the original evaluation observations rather than from the bootstrap distribution.

The fitted forecasting models, vulnerability-measurement parameters, predictor preprocessing transformations, and probability-recalibration functions are held fixed throughout resampling. Thus, the intervals describe uncertainty in evaluation performance conditional on the estimated forecasting system. They are not interpreted as incorporating the full uncertainty associated with generating the vulnerability outcome or selecting and fitting the model.

Table B5.1 reports country-clustered confidence intervals for the calibrated baseline breadth forecasts in the internal and external evaluation periods.

Table B5.1. Country-clustered uncertainty for the calibrated baseline breadth forecasts

Performance metric	Internal 2010–2017	External 2018–2023
ROC AUC	0.838 [0.800, 0.872]	0.819 [0.772, 0.861]
Average precision	0.595 [0.503, 0.681]	0.682 [0.576, 0.771]
Brier score	0.120 [0.101, 0.141]	0.141 [0.119, 0.161]
Expected calibration error	0.039 [0.026, 0.066]	0.057 [0.040, 0.088]
Precision	0.500 [0.422, 0.575]	0.538 [0.458, 0.618]
Recall	0.618 [0.524, 0.699]	0.675 [0.595, 0.752]
F1 score	0.553 [0.471, 0.623]	0.599 [0.522, 0.672]
False-alarm rate	0.161 [0.126, 0.198]	0.218 [0.166, 0.274]

Notes: Entries report point estimates followed by 95% country-clustered bootstrap confidence intervals in square brackets. Each interval is based on 1,000 bootstrap replications. Classification statistics use the 0.30 operating threshold. Expected calibration error is calculated using ten equal-width probability bins. Confidence intervals are conditional on the fitted forecasting pipeline and do not incorporate model-selection or outcome-measurement uncertainty.

The baseline model retains substantial discrimination in both evaluation periods. Internal ROC AUC is 0.838, with a 95% confidence interval of 0.800–0.872, while external ROC AUC is 0.819, with a corresponding interval of 0.772–0.861. Both intervals lie well above the no-discrimination reference of 0.5.

Average precision is 0.595 internally and 0.682 externally, with confidence intervals of 0.503–0.681 and 0.576–0.771, respectively. These values indicate substantial precision-recall performance relative to the corresponding event frequencies. However, because average precision depends on event prevalence, the higher external estimate should not be interpreted as evidence of superior temporal generalization without accounting for the different prevalence of vulnerability across periods.

Probability accuracy is somewhat weaker in the external period. The Brier score increases from 0.120 internally to 0.141 externally, while expected calibration error increases from 0.039 to 0.057.

The associated confidence intervals indicate sampling uncertainty around these estimates but do not establish a statistically significant deterioration between periods, since no paired test of the temporal difference is reported.

At the 0.30 operating threshold, external recall is 0.675 and precision is 0.538. The corresponding false-alarm rate is 0.218, with a confidence interval of 0.166–0.274. These results characterize uncertainty around the selected surveillance operating point rather than an institutionally optimal warning threshold. The country-clustered intervals support the conclusion that the selected model retains meaningful forecasting performance beyond the development sample. Whether its advantages over simpler forecasting rules are statistically distinguishable from zero is examined through paired model comparisons.

B.5.2 Paired model comparisons and incremental forecasting performance

The central inferential exercise evaluates whether the selected multivariate model improves on fitted lagged-score persistence when both models forecast the same observations. Comparing confidence intervals for separately estimated AUC or Brier statistics is insufficient for this purpose because the forecasts are evaluated against the same outcomes and their prediction errors are correlated.

For each vulnerability definition m, let $\mathcal{E}_m^{\text{pair}}$ denote the external-evaluation observations for which the selected model and fitted persistence benchmark both generate valid forecasts. The paired difference for performance measure q is defined as

$$\widehat{\Delta}_q^{(m)} = q\left(\widehat{\boldsymbol{p}}_{\text{selected}}^{(m)}, \boldsymbol{Y}^{(m)}\right) - q\left(\widehat{\boldsymbol{p}}_{\text{persistence}}^{(m)}, \boldsymbol{Y}^{(m)}\right),$$

where both performance statistics are evaluated over $\mathcal{E}_m^{\text{pair}}$.

Positive differences favour the selected model for ROC AUC and average precision. Conversely, negative differences in the Brier score favour the selected model because they indicate lower quadratic probability loss.

The paired bootstrap uses identical resampled country clusters for both models in each replication. For bootstrap draw b,

$$\widehat{\Delta}_q^{*(b,m)} = q_{\text{selected}}^{*(b,m)} - q_{\text{persistence}}^{*(b,m)}.$$

The corresponding 95% confidence interval is obtained from the empirical 2.5th and 97.5th percentiles of the 1,000 paired bootstrap differences.

This procedure preserves the dependence between the competing forecasts and prevents differences in observation coverage from affecting the estimated incremental performance.

Table B5.2 reports the paired external differences for the breadth, severity, and PCA-based vulnerability outcomes.

Table B5.2. Paired external forecasting differences relative to fitted lagged-score persistence, 2018–2023

Vulnerability definition	Δ AUC [95% CI]	Δ AP [95% CI]	Δ Brier [95% CI]
Breadth	+0.022 [0.006, 0.039]	+0.079 [0.050, 0.109]	−0.0086 [−0.0154, −0.0010]
Severity	+0.020 [0.004, 0.038]	+0.088 [0.044, 0.130]	−0.0255 [−0.0373, −0.0145]
PCA statistical component	−0.004 [−0.012, 0.004]	−0.008 [−0.025, 0.007]	+0.0013 [−0.0060, 0.0090]

Notes: Differences are calculated as selected multivariate model minus fitted lagged-score persistence on 718 matched country-year observations from 124 economies for each vulnerability definition. Confidence intervals are based on 1,000 paired country-clustered bootstrap replications. Positive AUC and AP differences favour the selected model, whereas negative Brier differences indicate better probability accuracy. The model architecture and vulnerability-measurement parameters remain fixed during bootstrap resampling.

For the baseline breadth measure, the selected model improves AUC by 0.022 relative to fitted persistence, with a 95% confidence interval of 0.006–0.039. The corresponding average-precision gain is 0.079, with an interval of 0.050–0.109. Both intervals exclude zero, indicating that the multivariate forecasting system contains incremental discrimination information beyond the previous breadth score.

The Brier difference is −0.0086, with a confidence interval of −0.0154 to −0.0010. This indicates improved probability accuracy, corresponding to the approximately 5.7% Brier skill reported in Section B.4. Although the reduction in probability loss is moderate, the paired interval excludes zero.

For the severity definition, the selected model similarly improves discrimination relative to fitted persistence. The AUC difference is +0.020, with a 95% confidence interval of 0.004–0.038, while average precision improves by 0.088, with an interval of 0.044–0.130.

The probability-accuracy improvement is more substantial under severity. The paired Brier difference is −0.0255, with a confidence interval of −0.0373 to −0.0145. This corresponds to an approximately 16.2% reduction in Brier loss relative to fitted persistence. The result indicates that the broader macro-fiscal predictor set contributes additional probability information when the

outcome incorporates the standardized magnitude of adverse threshold exceedances rather than simply counting their incidence.

The PCA-based statistical component produces a different pattern. The selected multivariate model has an AUC difference of −0.004 relative to persistence, with a confidence interval of −0.012 to 0.004. Average precision is also slightly lower, with a difference of −0.008 and an interval of −0.025 to 0.007. Neither interval excludes zero.

The paired Brier difference for the PCA outcome is +0.0013, with a confidence interval of −0.0060 to 0.0090. The positive point estimate indicates marginally higher probability loss for the selected model, but the confidence interval includes zero. Consequently, the evidence does not establish an incremental probability-forecasting advantage for the selected specification under the PCA-based definition.

These differences are substantively important. The PCA outcome is highly forecastable in absolute terms, but much of that predictability is already captured by its previous continuous score. In contrast, breadth- and severity-based vulnerability retain information that is not fully summarized by their respective lagged scores.

B.5.3 Comparison with conventional logistic regression

The selected model is also compared with conventional multivariate logistic regression, which uses the same eleven lagged macro-fiscal predictors but imposes a linear relationship in the log odds of vulnerability. This comparison addresses a different question from the persistence benchmark: whether the nonlinear forecasting specification adds predictive value relative to a conventional statistical model with the same predictor information. Table B5.3 reports the paired external differences for the baseline breadth outcome.

The selected model exceeds conventional logistic regression by approximately 0.020 AUC points, but the confidence interval includes zero. The estimated discrimination advantage is therefore not statistically distinguishable from zero at the conventional two-sided 5% level.

The average-precision improvement of 0.039 is more precisely estimated, with a confidence interval marginally above zero. This provides evidence of an incremental advantage in precision-recall performance, although the lower confidence bound is close to zero.

Table B5.3. Paired external forecasting differences relative to conventional logistic regression

Performance measure	Selected minus logistic	95% confidence interval
ROC AUC	+0.020	[−0.006, 0.048]
Average precision	+0.039	[0.0004, 0.080]
Brier score	−0.0074	[−0.0165, 0.0008]

Notes: Comparisons use 719 matched external-evaluation country-years from 124 economies and 1,000 paired country-clustered bootstrap replications. Positive AUC and AP differences and negative Brier differences favour the selected random forest. Confidence intervals are conditional on the fitted forecasting pipelines.

The Brier difference is −0.0074, indicating lower probability loss for the random forest, but its confidence interval also includes zero. Thus, the evidence does not support a general claim that the selected nonlinear model consistently dominates conventional logistic regression across all forecast-performance criteria.

This qualification is important because the model-selection exercise identifies several competitive forecasting specifications. The principal contribution of the analysis is not the superiority of a particular algorithm, but the assessment of whether broader predictive information provides incremental forecast skill under alternative constructions of fiscal vulnerability.

B.5.4 Interpretation and inferential scope

The country-clustered bootstrap results establish a distinction between absolute forecastability and incremental predictive information. The selected multivariate specification achieves substantial out-of-sample discrimination under all three principal vulnerability definitions, but its advantage over fitted persistence is not uniform.

For the breadth outcome, incremental gains are positive and statistically distinguishable from zero in AUC, average precision, and Brier loss. Under severity, the gains are also positive, with a particularly pronounced improvement in probability accuracy. Under the PCA-based definition, the corresponding differences are small and statistically indistinguishable from zero.

These results indicate that the measurement rule affects the extent to which the historical vulnerability score summarizes information about the subsequent vulnerability state. Under the PCA construction, lagged-score persistence captures most of the available predictive signal. Under breadth and severity, the broader lagged macro-fiscal information set contains additional information that is not fully represented by the preceding continuous score.

The interpretation is subject to two important qualifications. First, the reported confidence intervals are conditional on the fitted forecasting pipelines and do not incorporate model-selection

uncertainty or uncertainty arising from the estimation of vulnerability-measurement parameters. Second, the intervals characterize sampling variation within the observed external-evaluation regime; they do not establish that the same incremental gains would necessarily persist under future macroeconomic conditions.

Furthermore, confidence intervals excluding zero provide evidence of non-zero paired performance differences under the specified resampling procedure, but they do not by themselves establish the economic value of implementing a more complex forecasting system. That assessment also depends on calibration, surveillance operating thresholds, institutional warning costs, and whether the model identifies new vulnerability onsets rather than simply continuing states.

The evidence therefore supports a qualified conclusion: fiscal vulnerability is highly forecastable across alternative measurement constructions, but the incremental advantage of the selected multivariate model beyond persistence depends materially on how vulnerability is defined. Section B.6 examines whether these conclusions remain robust to alternative adverse thresholds, severity scaling, breadth cut-offs, and sample restrictions.

B.6 Robustness to Alternative Vulnerability Definitions and Sample Restrictions

The robustness analysis assesses whether the forecasting results depend on the construction of fiscal vulnerability or the availability of predictor observations. The selected depth-7 random-forest architecture and chronological evaluation protocol remain unchanged. Alternative outcomes are reconstructed using historical development-sample parameters, without repeating model selection. All results refer to the 2018–2023 external evaluation.

B.6.1 Alternative vulnerability definitions and thresholds

The analysis considers severity based on robust dispersion estimates, alternative component-tail thresholds of 20/80 and 30/70, and breadth classifications requiring at least two or four adverse dimensions instead of the baseline three. The PCA-based statistical component is retained as the third principal outcome definition.

Table B6.1. External forecasting performance under alternative vulnerability definitions

Vulnerability definition	AUC	AP	Brier	Brier skill vs. persistence (%)
Baseline breadth (≥3)	0.819	0.682	0.141	+5.7
Severity index	0.813	0.706	0.132	+16.2

Robust-scaled severity	0.873	0.713	0.093	+8.2
PCA statistical component	0.959	0.931	0.075	−1.8
Component thresholds 20/80	0.832	0.644	0.103	+9.5
Component thresholds 30/70	0.809	0.771	0.173	+5.0
Breadth threshold ≥2	0.816	0.824	0.178	+3.6
Breadth threshold ≥4	0.829	0.558	0.094	+2.2

Notes: AUC, AP, and Brier scores are calculated on 719 external country-year observations. Brier skill is calculated against fitted lagged-score persistence on 718 matched observations, as defined in Section B.4. Positive values indicate lower probability loss for the selected model. Differences in AP and Brier scores across outcome definitions partly reflect differences in event prevalence.

The selected model maintains substantial discrimination across the alternative constructions, with AUC values ranging from 0.809 to 0.873 for the breadth- and severity-based sensitivity specifications. Brier skill remains positive but varies from 2.2% to 16.2%. In contrast, the PCA-based outcome exhibits exceptionally high absolute discrimination without improvement over fitted persistence. These results reinforce the distinction between robust forecastability and measurement-dependent incremental predictive gains.

B.6.2 Leave-one-component-out robustness

To assess sensitivity to individual vulnerability dimensions, the breadth outcome is reconstructed after excluding each of the eight components in turn. The adverse thresholds and remaining outcome-construction parameters follow the historical development-sample protocol. The forecasting architecture is held fixed and re-estimated for each alternative outcome.

Table B6.2. Leave-one-component-out external forecasting performance

Excluded vulnerability component	**AUC**	**AP**	**Brier**
Government debt	0.795	0.611	0.138
Fiscal deficit	0.856	0.702	0.103
Government revenue	0.828	0.663	0.134
Government expenditure	0.846	0.640	0.103
Real GDP growth	0.861	0.671	0.108
Inflation	0.805	0.616	0.140
Current-account balance	0.803	0.607	0.127
Exchange-rate depreciation	0.814	0.654	0.137

Notes: Each specification excludes one dimension from the construction of the vulnerability outcome while retaining the locked forecasting architecture and eleven lagged predictors. All evaluations use 719 country-year observations over 2018–2023. Performance differences reflect changes in the constructed outcome and should not be interpreted as causal contributions of individual components.

External AUC ranges from 0.795 when government debt is excluded to 0.861 when real GDP growth is excluded. No single component's exclusion eliminates substantial predictive

discrimination. Nevertheless, the variation confirms that the composition of the vulnerability measure affects forecasting performance.

B.6.3 Predictor completeness and government-sector definitions

The baseline forecasting pipeline retains observations with missing lagged predictors through historical-sample median imputation. A supplementary complete-predictor exercise restricts estimation and evaluation to observations for which all eleven predictors are available. Additional robustness exercises replace the consolidated or harmonised government measures with general-government and central-government counterparts.

Table B6.3. Predictor-completeness and government-sector sensitivity

Specification	External observations	AUC	AP	Brier
Baseline consolidated/harmonised government	719	0.819	0.682	0.141
Complete-predictor sample	663	0.822	0.685	0.140
General-government definition	717	0.832	0.758	—
Central-government definition	224	0.835	0.501	—

Notes: The complete-predictor exercise requires all eleven lagged predictors to be observed and re-estimates the selected forecasting specification using the restricted historical sample. Alternative government-sector specifications differ in outcome construction and available country-year coverage. The central-government sample is substantially smaller than the baseline sample; its performance is therefore not directly comparable. Brier values for the government-sector alternatives are omitted from this condensed table.

Restricting the sample to complete predictor observations yields an AUC of 0.822, close to the baseline value of 0.819, indicating that baseline discrimination is not primarily driven by predictor imputation. General-government and central-government specifications also retain substantial discrimination, although differences in sample composition, particularly the smaller central-government evaluation sample, prevent direct interpretation of their relative performance as improvements.

The robustness exercises support the main finding: fiscal vulnerability remains forecastable under alternative measurement and sample specifications, while the incremental benefit of multivariate forecasting beyond persistence depends on the definition of the vulnerability outcome.

B.7 Vulnerability Onset and Continuation

Aggregate forecasting performance may reflect the persistence of existing vulnerability rather than the ability to identify newly emerging adverse conditions. To distinguish these cases, the external evaluation is partitioned according to the preceding year's vulnerability state. For each definition m, onset and continuation observations are defined as

$$\text{Onset: } Y_{i,t-1}^{(m)} = 0, \qquad \text{Continuation: } Y_{i,t-1}^{(m)} = 1.$$

The same externally generated predictions are evaluated separately within each subsample. No transition-specific models are estimated. Incremental performance is assessed against fitted lagged-score persistence using matched observations and 1,000 country-clustered bootstrap replications.

B.7.1 Forecasting vulnerability onset

Table B7.1 reports the external forecasting performance for vulnerability onset under the three principal outcome definitions, including paired differences relative to fitted lagged-score persistence and country-clustered confidence intervals.

Table B7.1. External forecasting performance for vulnerability onset, 2018–2023

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

Notes: Onset observations are country-years that were non-vulnerable in the preceding year. Events represent transitions into vulnerability in the current year. AUC and AP refer to the selected random forest. Differences are measured against fitted lagged-score persistence on matched observations. Confidence intervals use country-clustered bootstrap resampling.

Under the breadth definition, the selected model achieves an onset AUC of 0.721 and AP of 0.338. Its incremental advantages over fitted persistence are +0.043 AUC points and +0.098 AP points, with both confidence intervals excluding zero. The multivariate forecasting system therefore contains information about new breadth-vulnerability episodes beyond that captured by the preceding breadth score.

The corresponding gains are weaker under severity and absent in AUC under PCA. Although the PCA-based onset AUC is relatively high at 0.875, its incremental discrimination advantage over persistence is effectively zero. High onset predictability therefore does not necessarily imply additional early-warning information from the broader predictor set.

B.7.2 Forecasting vulnerability continuation

Continuation observations correspond to country-years already classified as vulnerable in the preceding year. Performance within this subsample assesses the ability to distinguish continued

vulnerability from subsequent exits. Table B7.2 compares the discriminatory performance of the selected forecasting model for vulnerability onset and continuation across the three definitions.

Table B7.2. External discrimination for vulnerability onset and continuation

Vulnerability definition	Onset AUC	Continuation AUC
Breadth	0.721	0.779
Severity	0.663	0.829
PCA statistical component	0.875	0.814

Notes: The same externally generated model probabilities are evaluated separately conditional on the preceding vulnerability state. Onset and continuation are distinct evaluation subsamples; differences between their AUC values are descriptive and are not formal paired comparisons.

Continuation discrimination exceeds onset discrimination for breadth and severity, particularly under the severity definition, where AUC increases from 0.663 to 0.829. This suggests that the strong aggregate performance under severity is associated more closely with identifying the continuation or resolution of existing vulnerability than with anticipating new entries.

For PCA, onset AUC exceeds continuation AUC, although neither comparison establishes incremental forecast information beyond persistence.

The onset analysis provides the clearest evidence of incremental early-warning content under the breadth definition. Under severity, the model remains useful for monitoring existing vulnerability, while the PCA-based outcome is highly predictable largely through information already contained in its lagged score. These findings reinforce the measurement dependence of the forecasting system's early-warning value.

B.8 Operating-Threshold and Supplementary Forecasting Diagnostics

This section examines the sensitivity of forecast classifications to the surveillance threshold, compares the locked forecasting specifications across model families, and reports supplementary performance and predictor-importance diagnostics. All analyses retain the chronological evaluation design and baseline breadth-vulnerability definition.

B.8.1 Sensitivity to the surveillance operating threshold

The baseline analysis uses a probability threshold of 0.30 to classify vulnerability warnings. Because the appropriate threshold depends on the relative costs of missed vulnerabilities and false alarms, alternative thresholds are evaluated without re-estimating the forecasting model. Table B8.1 summarizes external classification performance at selected thresholds.

Table B8.1. External warning-threshold sensitivity, 2018–2023

Threshold	Precision	Recall	F1	False-alarm rate	Missed-vulnerability rate
0.10	0.399	0.914	0.556	0.519	0.086
0.20	0.456	0.766	0.572	0.345	0.234
0.30	0.538	0.675	0.599	0.218	0.325
0.40	0.600	0.609	0.605	0.153	0.391
0.50	0.694	0.518	0.593	0.086	0.482

Notes: All statistics use the same 719 external evaluation observations and calibrated probabilities. The 0.30 threshold is the prespecified operating point. The false-alarm rate is the proportion of non-vulnerable observations incorrectly classified as vulnerable, while the missed-vulnerability rate is the proportion of vulnerable observations incorrectly classified as non-vulnerable.

Lower thresholds increase recall but generate more false alarms. At 0.30, recall is 0.675 and the false-alarm rate is 0.218. Raising the threshold to 0.50 reduces false alarms to 0.086 but increases missed vulnerabilities to 0.482. Figure B8.1 illustrates the associated trade-offs across the full evaluated threshold range. No threshold is interpreted as institutionally optimal without an explicit warning-cost function.

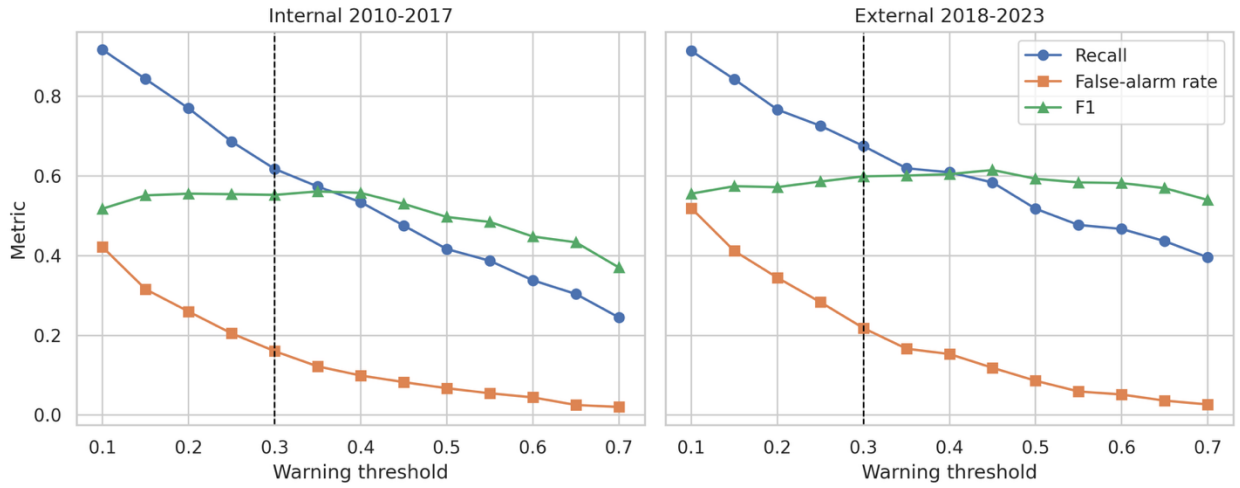


Figure B8.1. Recall, false-alarm rate, and F1 across warning thresholds

Notes: The vertical reference indicates the prespecified 0.30 threshold. Curves are calculated from the calibrated external forecasts.

B.8.2 External performance across supervised-learning families

To examine whether the results depend on the selected random-forest architecture, the development-period configurations from alternative model families are evaluated using the locked external forecasting protocol. Model selection is not reopened. Table B8.2 reports the leading external comparisons.

Table B8.2. External discrimination across leading supervised-learning families

Model family	AUC	AP
XGBoost	0.825	0.693
Histogram gradient boosting	0.823	0.668
Selected random forest	0.819	0.682
Gradient boosting	0.818	0.683
AdaBoost	0.815	0.667
Neural network	0.813	0.671
Extra Trees	0.810	0.646
Logistic regression	0.799	0.643

Notes: Results refer to 719 observations over 2018–2023. All models follow the common historical estimation, recalibration, and external evaluation protocol. The random forest remains the prespecified selected specification irrespective of alternative models' external performance.

External AUC varies relatively little among the leading ensemble methods. XGBoost achieves 0.825 compared with 0.819 for the selected random forest. These small ex post differences do not justify revising the development-period model-selection decision. Figure B8.2 provides the corresponding model-family comparison.

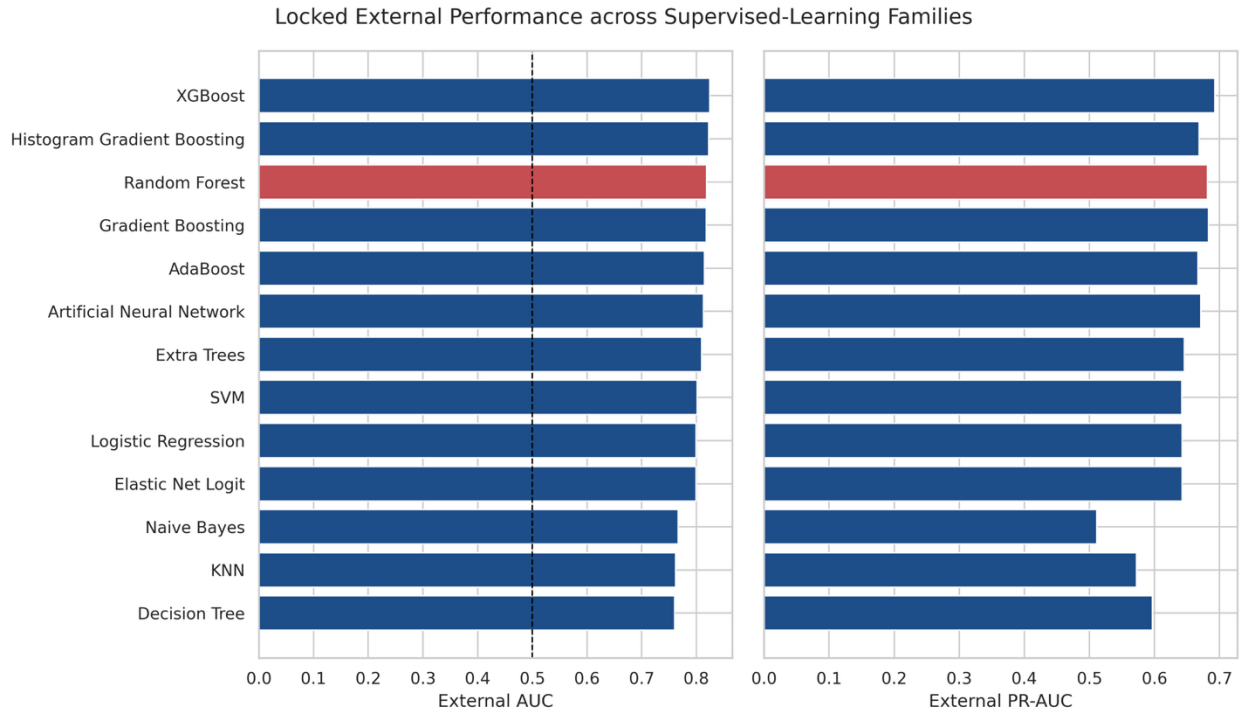


Figure B8.2. External discrimination across supervised-learning families

B.8.3 Income-group forecasting performance

Forecast performance is also assessed across the three static GMD income groups. The same pooled forecasting model is evaluated within each subgroup; no income-specific models are estimated. Table B8.3 reports external performance by income group.

Table B8.3. External forecasting performance by income group

Income group	Observations	AUC	AP	Brier
Low income	123	0.880	0.835	0.114
Lower-middle income	286	0.846	0.650	0.132
Upper-middle income	310	0.770	0.625	0.161

Notes: Income groups follow the static classifications in the June 2026 GMD release. Differences are descriptive and reflect variations in sample composition and vulnerability incidence.

Discrimination is highest among low-income economies and lowest among upper-middle-income economies. The corresponding Brier scores indicate weaker probability accuracy in the latter group. These differences demonstrate heterogeneity in forecasting performance but do not establish that separate income-group models would improve predictive accuracy.

B.8.4 Predictor-importance diagnostics

Permutation importance and SHAP diagnostics are used to describe the information employed by the selected forecasting model. Permutation importance measures the decline in ROC AUC when a predictor is randomly permuted, while SHAP values summarize contributions to the fitted prediction function. Figure B8.3 reports permutation importance, and Figure B8.4 presents global SHAP importance.

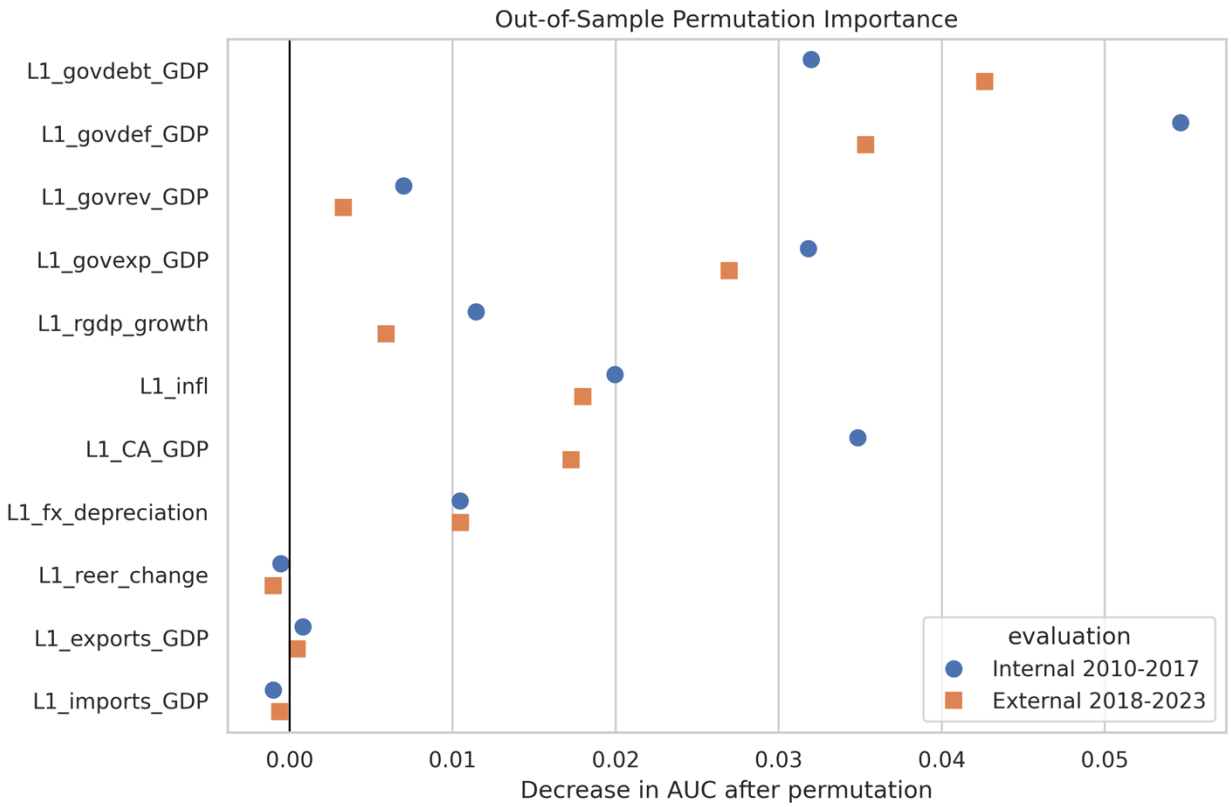


Figure B8.3. Out-of-sample permutation importance

Notes: Importance is measured by the decrease in ROC AUC following predictor permutation in the internal and external evaluation samples.

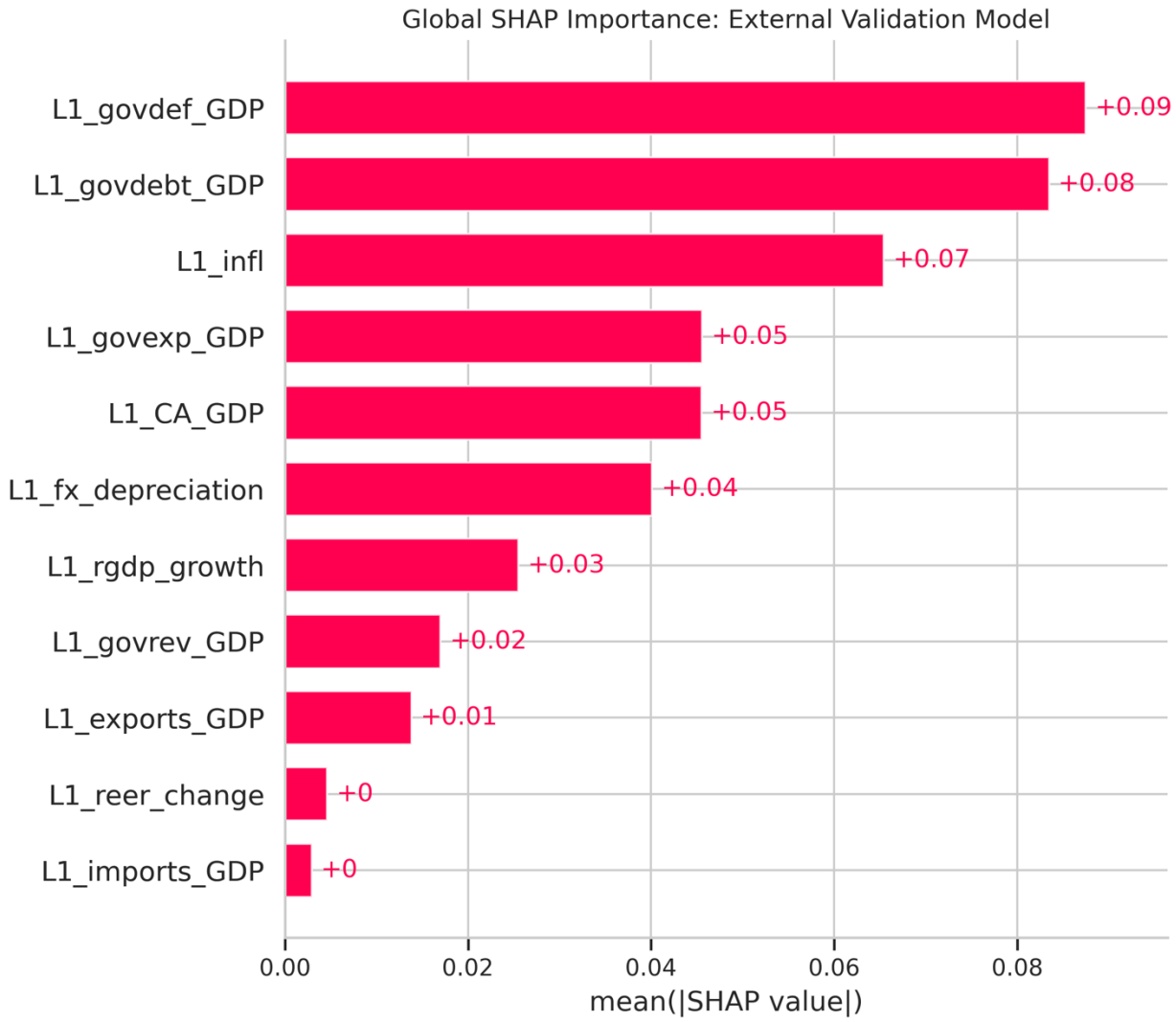


Figure B8.4. Global SHAP predictor importance

Notes: Bars represent mean absolute SHAP contributions for the external forecasting model. Contributions describe the fitted prediction rule and are not causal effects.

The two approaches identify lagged fiscal balance and government debt as important predictors, alongside government expenditure, inflation, and current-account conditions. Their relative importance differs somewhat across methods because permutation importance measures dependence of predictive discrimination on a variable, whereas SHAP measures contributions to individual predictions.

These diagnostics support the relevance of the broader macro-fiscal information set without assigning causal interpretations to predictor contributions.

B.9 Country-Level Surveillance Forecasts and Cross-Definition Ranking Agreement

Country-level forecasts are obtained from the pooled forecasting models for the 2018-2023 external evaluation period. For each economy, predicted probabilities are averaged across available years separately for breadth, severity, and PCA-based vulnerability. Rankings are based on the mean breadth probability. Annual ranking agreement is assessed using Spearman correlation, Kendall's \(\tau\), and overlap in the highest predicted-probability quintile.

Table B9.1. Agreement in country-level surveillance rankings, 2018–2023

Forecast pair	Mean Spearman	Annual range	Mean Kendall τ	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. Top-20% overlap measures agreement in membership of the highest predicted-probability quintile.

Table B9.1 shows that rankings are most closely aligned between breadth and severity, while agreement with PCA is weaker. Even breadth and severity share only 60.3% of their upper-quintile memberships on average. Table B9.2 provides the corresponding country-level mean forecasts for all 124 economies, ordered by baseline breadth ranking.

Table B9.2. Country-level mean predicted fiscal-vulnerability probabilities, 2018–2023
B = breadth, S = severity, P = PCA. N = years observed. Values are probabilities, not percentages.

Rank	Economy	N	B	S	P
1	Yemen	5	0.935	0.423	0.232
2	Sudan	5	0.919	0.829	0.044
3	Lebanon	6	0.879	0.922	0.456
4	Dominica	6	0.856	0.953	0.992
5	Argentina	6	0.833	0.160	0.897
6	Mozambique	6	0.748	0.676	0.805
7	Suriname	6	0.704	0.337	0.522
8	Bhutan	6	0.692	0.707	0.749
9	South Sudan	6	0.677	0.633	0.845
10	Maldives	6	0.665	0.612	0.855
11	Brazil	6	0.616	0.167	0.946
12	Sri Lanka	6	0.611	0.201	0.012
13	Montenegro	6	0.604	0.748	0.991
14	Timor-Leste	6	0.603	0.764	0.903
15	Senegal	6	0.592	0.360	0.541

Rank	Economy	N	B	S	P
16	Egypt	6	0.566	0.092	0.119
17	Algeria	6	0.536	0.262	0.865
18	Liberia	6	0.523	0.358	0.626
19	Fiji	6	0.483	0.279	0.704
20	Jordan	6	0.483	0.102	0.446
21	Zambia	6	0.480	0.190	0.354
22	São Tomé and Príncipe	6	0.478	0.549	0.345
23	Cabo Verde	6	0.475	0.350	0.362
24	Lesotho	6	0.453	0.866	0.986
25	Tunisia	6	0.453	0.122	0.787
26	St Vincent and the Grenadines	6	0.448	0.404	0.533
27	Eritrea	2	0.444	0.649	0.498
28	Bolivia	6	0.422	0.110	0.859
29	Mongolia	6	0.420	0.242	0.793
30	Tuvalu	6	0.415	0.960	0.948
31	Haiti	6	0.406	0.089	0.001
32	Ghana	6	0.403	0.140	0.134
33	Burundi	6	0.400	0.282	0.271
34	Pakistan	6	0.400	0.059	0.001
35	Namibia	6	0.399	0.182	0.957
36	Sierra Leone	6	0.398	0.087	0.001
37	Türkiye	6	0.389	0.055	0.551
38	Kiribati	6	0.387	0.946	0.924
39	Malawi	6	0.375	0.414	0.060
40	Iran	6	0.370	0.089	0.002
41	Mauritius	6	0.354	0.147	0.290
42	Nigeria	6	0.335	0.073	0.001
43	Angola	6	0.328	0.091	0.005
44	Rwanda	6	0.319	0.156	0.374
45	India	6	0.313	0.085	0.147
46	Grenada	6	0.301	0.271	0.468
47	The Gambia	6	0.294	0.044	0.011
48	El Salvador	6	0.291	0.072	0.370
49	Moldova	6	0.285	0.234	0.875
50	Ukraine	6	0.269	0.263	0.966
51	Colombia	6	0.263	0.059	0.567
52	Belize	6	0.254	0.079	0.184

Rank	Economy	N	B	S	P
53	Jamaica	6	0.245	0.059	0.166
54	Marshall Islands	6	0.236	0.933	0.913
55	Tonga	6	0.235	0.309	0.977
56	Botswana	6	0.234	0.115	0.644
57	Vanuatu	6	0.234	0.252	0.940
58	Laos	6	0.233	0.098	0.002
59	Micronesia	1	0.231	0.904	0.920
60	Saint Lucia	6	0.228	0.191	0.164
61	North Macedonia	6	0.216	0.070	0.588
62	Comoros	6	0.216	0.129	0.011
63	Georgia	6	0.210	0.116	0.328
64	South Africa	6	0.203	0.066	0.496
65	Kyrgyzstan	6	0.200	0.202	0.865
66	Serbia	6	0.200	0.172	0.971
67	DR Congo	6	0.199	0.038	0.001
68	Republic of the Congo	4	0.199	0.112	0.014
69	Uganda	6	0.196	0.045	0.001
70	Kenya	6	0.195	0.049	0.009
71	Iraq	6	0.192	0.116	0.805
72	Solomon Islands	6	0.192	0.167	0.939
73	Burkina Faso	6	0.190	0.068	0.154
74	China	6	0.171	0.041	0.495
75	Niger	6	0.171	0.233	0.015
76	Albania	6	0.165	0.056	0.560
77	Ecuador	6	0.165	0.088	0.936
78	Afghanistan	6	0.159	0.498	0.165
79	Guinea-Bissau	6	0.147	0.041	0.006
80	Somalia	2	0.146	0.108	0.002
81	Samoa	6	0.144	0.122	0.698
82	Bosnia and Herzegovina	6	0.144	0.163	0.961
83	Eswatini	6	0.142	0.050	0.316
84	Myanmar	5	0.132	0.087	0.004
85	Morocco	6	0.132	0.061	0.180
86	Belarus	6	0.126	0.075	0.923
87	Nepal	6	0.120	0.095	0.043
88	Kosovo	6	0.119	0.067	0.338
89	Uzbekistan	6	0.117	0.036	0.083

Rank	Economy	N	B	S	P
90	Bangladesh	6	0.111	0.037	0.000
91	Benin	6	0.099	0.032	0.001
92	Côte d'Ivoire	6	0.097	0.033	0.001
93	Madagascar	6	0.095	0.048	0.001
94	Azerbaijan	6	0.093	0.054	0.672
95	Chad	5	0.088	0.050	0.000
96	Gabon	4	0.087	0.044	0.001
97	Papua New Guinea	6	0.086	0.063	0.005
98	Tajikistan	6	0.086	0.040	0.272
99	Dominican Republic	6	0.078	0.049	0.001
100	Indonesia	6	0.077	0.036	0.000
101	Armenia	6	0.074	0.055	0.119
102	Kazakhstan	6	0.073	0.039	0.004
103	Mali	6	0.073	0.047	0.010
104	Philippines	6	0.072	0.055	0.012
105	Palestine	6	0.070	0.190	0.022
106	Guatemala	6	0.070	0.035	0.000
107	Cambodia	6	0.064	0.175	0.008
108	Peru	6	0.063	0.055	0.006
109	Togo	6	0.061	0.034	0.004
110	Djibouti	6	0.057	0.050	0.134
111	Cameroon	6	0.054	0.032	0.000
112	Equatorial Guinea	6	0.054	0.113	0.006
113	Mauritania	6	0.052	0.148	0.006
114	Thailand	6	0.051	0.051	0.006
115	Paraguay	6	0.049	0.042	0.001
116	Tanzania	6	0.048	0.032	0.000
117	Turkmenistan	6	0.047	0.082	0.003
118	Malaysia	6	0.046	0.051	0.008
119	Nicaragua	6	0.044	0.064	0.133
120	Mexico	6	0.042	0.052	0.027
121	Central African Republic	2	0.040	0.032	0.001
122	Guinea	6	0.038	0.030	0.001
123	Honduras	6	0.037	0.057	0.065
124	Vietnam	6	0.032	0.036	0.003

Notes: Probabilities are arithmetic means of calibrated country-year forecasts over available years in 2018–2023. Countries are ordered by descending mean breadth probability. N indicates the number of observed years, which varies because the panel is unbalanced. The results are pooled-model surveillance forecasts, not

Rank	Economy	N	B	S	P

estimates from country-specific models, sovereign-default probabilities, or formal risk ratings. Comparisons involving countries with short histories require particular caution.

Table B9.2 shows considerable heterogeneity across economies and outcome definitions. Yemen, Sudan, Lebanon, and Dominica record the highest mean breadth probabilities. However, the relative ordering changes substantially under severity and PCA. This confirms that the definition of fiscal vulnerability affects which economies receive greater surveillance attention, even when all three models are based on the same underlying macro-fiscal information set.

B.10 Post-Evaluation Sensitivity Analysis: 2024

The 2024 observations are evaluated separately from the principal external period to assess forecasting performance on the latest available analytical observations. The sample comprises 107 economies and 29 breadth-vulnerability events. The fixed development-sample vulnerability definition and prespecified 0.30 warning threshold are retained.

Table B10.1. Post-evaluation forecasting performance, 2024

Performance metric	2024 sensitivity
Observations	107
Vulnerability events	29
Event share	0.271
ROC AUC	0.890
Average precision	0.773
Brier score	0.111
Log loss	0.361
Expected calibration error	0.096
Precision	0.605
Recall	0.793
F1 score	0.687
Balanced accuracy	0.800
False-alarm rate	0.192
Missed-vulnerability rate	0.207

Notes: Results are obtained from the separate 2024 sensitivity predictions. Classification metrics use the 0.30 operating threshold. The 2024 observations are excluded from the principal external evaluation covering 2018–2023.

As reported in Table B10.1, the 2024 forecast achieves an AUC of 0.890, average precision of 0.773, and Brier score of 0.111. At the 0.30 threshold, recall reaches 0.793, with a false-alarm rate of 0.192. These results indicate substantial predictive discrimination in the additional evaluation

year. However, expected calibration error of 0.096 indicates remaining probability-calibration discrepancies. Given the single-year horizon and smaller sample, the results are interpreted as supplementary evidence rather than a replacement for the principal multi-year external evaluation.