GRAPE Working Paper #114

When 3% Means Nothing: Calibrating Escalation Limits to a Bank's Own Forecasting Error Distribution

Marcin Dec

FAME | GRAPE, 2026

Foundation of Admirers and Mavens of Economics
Group for Research in Applied Economics

When 3% Means Nothing: Calibrating Escalation Limits to a Bank's Own Forecasting Error Distribution

Marcin Dec
FAME|GRAPE

Abstract

Forecasting accuracy for Net Interest Income (NII) and Interest Rate Risk in the Banking Book (IRRBB) is central to banks' earnings stability, balance-sheet management, and supervisory credibility. Yet many institutions continue to apply fixed deviation thresholds (for example, 3/4/5%) to govern forecast performance, even though forecast uncertainty widens with the horizon and may exhibit heavy-tailed behavior. Such limits therefore lack a consistent probabilistic interpretation and often misalign with the statistical properties of the underlying forecasting process. This paper develops an integrated, probability-coherent framework for monitoring NII forecast errors and assessing IRRBB limit breaches. First, drawing on the Federal Reserve's use of Root Mean Squared Error (RMSE) and fan charts to communicate forecast uncertainty, we construct horizon-specific, quantile-anchored thresholds that preserve consistent meaning across forecast horizons. The framework incorporates interval-forecast evaluation (unconditional and conditional coverage tests), quantile elicitability, bias-dispersion decomposition, and extreme-value modeling of rare outcomes. Second, we extend the methodology to IRRBB by quantifying the probability that limits on changes in NII (ΔNII) are breached solely due to forecast or model uncertainty.

Keywords:
Net Interest Income forecasting, Interest rate risk in the banking book (IRRBB), Quantile-based escalation thresholds, Forecast uncertainty.

JEL Classification:
G21, G32, C58, G28

Corresponding author: :
Marcin Dec (m.dec@grape.org.pl).

Acknowledgments
The financial support of National Science Centre (grant UMO-2020/37/N/HS4/02202) is gratefully acknowledged.

Published by: FAME | GRAPE
ISSN: 2544-2473

1. Introduction

I'd rather be vaguely right, than precisely wrong
John Maynard Keynes

Forecasts of Net Interest Income (NII) play a central role in banks' earnings planning, Interest Rate Risk in the Banking Book (IRRBB) management, balance-sheet steering, and capital strategy. Supervisory expectations - particularly those articulated in the Federal Reserve and Office of the Comptroller of the Currency's Supervisory Guidance on Model Risk Management (SR 11-7) and OCC Appendix D: Heightened Standards - require banks to maintain robust processes for ongoing outcomes analysis, clear escalation logic, and risk-appetite frameworks in which limits are grounded in empirical evidence rather than static heuristics. Yet in practice, many institutions still rely on fixed percentage deviation thresholds (for example, 3%, 4%, or 5%) to govern forecast performance. Although operationally simple, such thresholds lack a stable statistical meaning across horizons because forecast uncertainty widens and can become asymmetric or heavy-tailed as the horizon extends.

A natural benchmark for horizon-specific uncertainty comes from the Federal Reserve's Summary of Economic Projections (SEP). Since 2007, the Federal Open Market Committee (FOMC) has published "average historical projection error ranges," defined as Root Mean Squared Errors (RMSEs) of major forecasters over the preceding twenty years. Reifschneider and Tulip (2017) show that ±1 RMSE approximates a 70% prediction interval under unbiased, symmetric errors, and that this mapping is empirically robust across macroeconomic variables, including short-rate forecasts. Their results also highlight several properties essential for bank-level NII forecasting: (i) macro forecasts exhibit substantial dispersion, (ii) different forecasters have broadly similar accuracy, (iii) uncertainty about real activity and interest rates increased after the financial crisis, and (iv) fan charts based on RMSE intervals provide a reasonable approximation to future uncertainty under typical conditions. This structured RMSE-to-coverage mapping provides a clear precedent for designing horizon-specific NII tolerance and escalation thresholds, and it emphasizes why fixed limits (e.g., 3%) cannot have consistent risk meaning across 3, 6, 12, or 24-month horizons.

The statistical evaluation of NII forecast performance connects directly to the literature on interval-forecast evaluation. Christoffersen (1998) develops a theoretically complete framework for evaluating prediction intervals under both unconditional coverage - i.e., whether the long-run frequency of exceedances matches the nominal level - and conditional coverage - i.e., whether exceedances are independent and not clustered, as would be expected under a well-specified model. These conditional dynamics matter for banking-book forecasting because model changes, scenario processes, and behavioral parameters can induce periods of instability or serial dependence in errors. Kupiec's Proportion of Failures (POF) test (1995) provides a simpler unconditional benchmark, widely adopted in the 1996 Basel "traffic-light" framework for back-testing Value-at-Risk (VaR), where the number of exceedances is mapped into green/yellow/red supervisory zones. This structure illustrates how probabilistic coverage targets can be operationalized into governance tiers.

Recent advances in forecast evaluation emphasize elicitability - the property that a statistical functional (such as a quantile or prediction interval) can be uniquely identified by a proper scoring rule. Fissler, Frongillo, Hlavinová, and Rudloff (2021) demonstrate that quantiles and set-valued prediction intervals are jointly elicitable, providing firm theoretical support for designing escalation thresholds using quantiles (e.g., 80%, 90%, 97.5%) rather than arbitrary percentages. Their results imply that quantile-based thresholds yield consistent decision rules under model comparison and can be formally evaluated using appropriate scoring functions.

Complementary strands of literature address differences in predictive accuracy. Diebold and Mariano (1995) propose a test for comparing predictive ability across models without assuming homoskedastic, independent forecast errors - an essential property for evaluating alternative NII forecasting methodologies. Giacomini and White (2006) extend this line of work with a Conditional Predictive Ability (CPA) test, enabling fair comparison under realistic estimation uncertainty and dynamic environments. These tools support model-risk management requirements for robust outcomes analysis and justify empirically selecting among alternative escalation frameworks based on predictive performance.

Because NII errors and long-horizon IRRBB outcomes may exhibit heavy tails, the framework draws on Extreme Value Theory (EVT). The Pickands–Balkema–de Haan (PBdH) theorem shows that the distribution of excesses over a high threshold converges to a Generalized Pareto Distribution (GPD), forming the basis of the Peaks-Over-Threshold (POT) approach. This limit result justifies using POT–GPD modeling to stabilize escalation tiers for rare but severe forecasting misses, analogous to extensions of the VaR "traffic-light" backtest to the Expected Shortfall context.

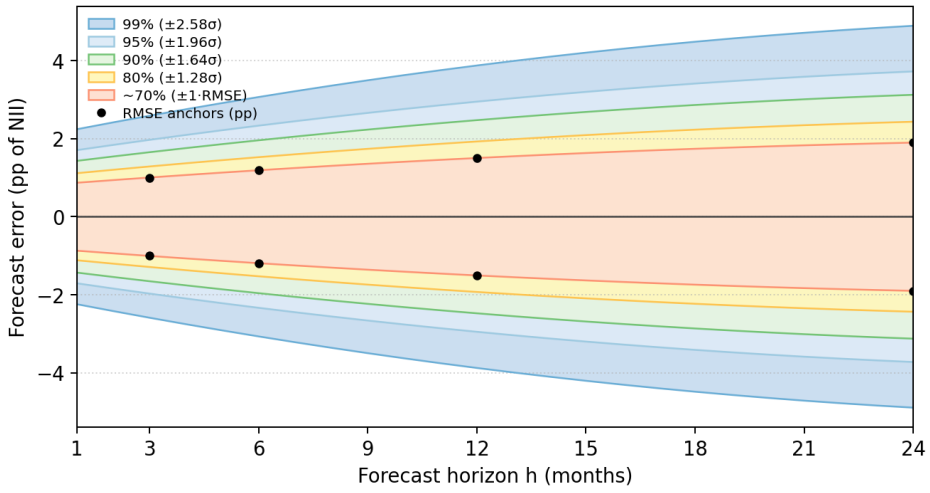
The regulatory environment for IRRBB further motivates a probability-coherent treatment. The Basel Committee's IRRBB Standard (BCBS 368, 2016) specifies six standardized interest-rate shock scenarios, behavioral modeling constraints (especially for non-maturity deposits, NMDs), currency-level granularity. This - together with enhanced disclosure expectations - suggest that limit monitoring should reflect both scenario structure and parameter uncertainty. They also imply that governance frameworks should quantify how likely it is that ΔNII limits are breached purely due to forecast or modeling uncertainty, an aspect generally absent from existing academic and practitioner literature.

Taken together, these strands highlight that a modern NII/IRRBB monitoring framework should (i) express limits as probability-calibrated quantiles, (ii) validate them via coverage statistics, (iii) incorporate bias-dispersion decomposition and tail-risk modeling, and (iv) link these components to regulatory shock structures and risk-appetite governance. This paper operationalizes these ideas, extends them to IRRBB limit breach probabilities, and provides a unified architecture grounded in both economic forecasting research and prudential standards.

2. Setting Probabilistically Coherent Limits

We frame the monitoring problem around horizon-specific absolute forecast errors and anchor escalation thresholds to quantiles, so that each escalation tier (C1, C2, C3) carries a stable probabilistic meaning regardless of the forecast horizon h. For communication and governance, we exploit the Federal Reserve's Summary of Economic Projections (SEP) practice of mapping Root Mean Squared Error (RMSE) to coverage - where, under unbiased and symmetric errors, a ±1 RMSE interval approximates a 70% prediction band - while in production we estimate limits from the bank's empirical error distribution using the empirical distribution function and kernel density estimation. This dual approach preserves interpretability for non-technical stakeholders and fidelity to institution-specific data.

Figure 1. RMSE-base prediction bands over 24 monthly horizons (illustrative)



Formally, let A_{t+h} denote realized Net Interest Income (NII) at horizon h, and $F_t^{(h)}$ the corresponding forecast made at time t. We measure scale-free relative forecast errors as the ratio of forecast miss to realized NII:

$$e_{t,h} = \frac{A_{t+h} - F_t^{(h)}}{A_{t+h}}$$

Escalation thresholds may be defined as horizon-specific quantiles of the absolute error distribution, ensuring that each committee tier has a consistent probabilistic interpretation:

$$L_h^{(k)} = Q_h(\alpha_k)$$

To provide a transparent communication framework, we apply the Federal Reserve's RMSE-to-coverage mapping, where symmetric, unbiased forecast errors produce quantile levels equal to sigma-multiples of the RMSE, i.e. in:

$$Q_h(\alpha) = \sigma_h \cdot \Phi^{-1}\left(\frac{\alpha + 1}{2}\right)$$

which maps each coverage level α to a "sigma-multiple" of the horizon-specific dispersion σ_h. Throughout, σ_h denotes the horizon-specific dispersion of forecast errors and under the SEP-style Normal/RMSE mapping used for communication and corresponds exactly to the RMSE at that horizon. So that, for example, 80%, 90%, 95%, and 99% two-sided limits correspond to 0.84, 1.28, 1.64, and 2.33 standard deviations, respectively. This mirrors the SEP fan-chart construction and provides a clear, horizon-aware language for risk appetite and escalation.

To support governance dashboards, we construct a standardized exceedance score that expresses each realized error as a proportion of the 95% boundary, enabling uniform interpretation of signal strength across horizons:

$$S_{t,h} = \frac{|e_{t,h}|}{Q_h(0.95)}$$

so that $S_{t,h} = 1$ lies exactly on the 95% contour and values above one indicate tail realizations relative to the intended C2 band.

Because persistent estimation bias can contaminate dispersion estimates and distort governance signals, we should separate systematic bias from random dispersion. This approach aligns with SR 11-7 expectations for outcomes analysis and avoids normalizing structural model bias into risk-appetite tolerances:

$$\tilde{e}_{t,h} = e_{t,h} - \mu_h$$

where: μ_h stands for the sample average error for a given forecasting horizon h.

Since long-horizon forecast errors can be heavy-tailed and samples in the deep tail are sparse, the empirical quantile Q_h usually becomes unstable exactly where governance needs the most reliability. To restore stability while preserving statistical fidelity to the bank's data, we could calibrate the outermost escalation tier with a Peaks-Over-Threshold construction applied to the absolute, bias-adjusted horizon-h errors. Notice that $\tilde{e}_{t,h}$ removes the horizon-specific mean to keep structural misspecification visible in bias logs rather than absorbed into dispersion. Write $X_{t,h} = |\tilde{e}_{t,h}|$ and fix a high threshold u_h for horizon h. In practice, u_h is chosen by combining statistical diagnostics - such as stability of the fitted tail-shape parameter as the threshold varies around high percentiles - and governance pragmatism that targets a high but data-sustaining level, for example the 90th to 95th empirical percentile of $\{X_{t,h}\}$ at horizon h. This choice delivers enough exceedances to estimate a parsimonious tail model while remaining sufficiently far into the tail for the asymptotic approximation to hold. Under the Pickands–Balkema–de Haan limit result, the exceedances $Y_{t,h} = X_{t,h} - u_h$ conditional on $X_{t,h} > u_h$ are well approximated by a Generalized Pareto Distribution, $Y_{t,h} \sim \text{GPD}(\xi_h, \beta_h)$, with shape ξ_h and scale β_h. Parameters (ξ_h, β_h) are estimated at each horizon by maximum likelihood or probability-weighted moments, and their stability may be checked across nearby thresholds together with standard tail QQ-plots. A positive $\xi_h$indicates a heavy tail, whereas $\xi_h \approx 0$ an exponential-like tail. Denote by $p_h = \mathbb{P}(X_{t,h} > u_h)$ the empirical exceedance rate at horizon h, i.e., the fraction of observations above u_h. For a desired deep-tier coverage level α (for example, the 97.5th or 99th percentile on a two-sided absolute-error scale), the corresponding POT–GPD quantile for the full distribution of $X_{t,h}$ is obtained by inverting the tail model and splicing it back at u_h. The resulting escalation threshold is:

$$L_h^{(\text{EVT})} = u_h + \frac{\beta_h}{\xi_h}\left[\left(\frac{1-\alpha}{p_h}\right)^{-\xi_h} - 1\right]$$

with the exponential-tail limit $L_h^{(\text{EVT})} = u_h + \beta_h \ln\left(p_h/(1-\alpha)\right)$ when $\xi_h \to 0$. This $L_h^{(\text{EVT})}$replaces the raw empirical quantile only for the outermost tier. The interior tiers continue to use $Q_h(\alpha)$ estimated from the empirical distribution function or kernel-smoothed EDF to preserve fidelity where data are rich. Escalation tiers must be materially distinct. In practice, the EVT overlay should used only for the outer tier to avoid over-fitting and to ensure that ordinary governance remains driven by data-rich central quantiles.

To prevent tier cascades - where a small overrun of the first tier mechanically breaches subsequent tiers - we could impose a minimum spacing rule (spacing discipline) proportional to horizon-specific dispersion, ensuring structural separation even under fat-tailed error distributions:

$$L_h^{(k+1)} - L_h^{(k)} \geq c \cdot \sigma_h$$

with α-levels are – for instance - set to $(0.8, 0.9)$ and $c = 0.5$. This preserves separation between tiers even under thicker tails and echoes the spirit of traffic-light separation familiar from risk back-testing. As a result, the calibrated multi-tier and multi horizon limit set could be calculated as:

$$\mathcal{L} = \left(L_h^{(1)},\ L_h^{(2)},\ L_h^{(EVT)}\right), h \in [1,2, \dots 24]$$

where $L_h^{(2)} \leq L_h^{(EVT)} - c \cdot \sigma_h$, and $L_h^{(1)} \leq L_h^{(2)} - c \cdot \sigma_h$. This way, we keep committee actions materially distinct across levels, while quarterly unconditional and conditional coverage tests verify that realized deep-tier exception rates align with design targets and that exceptions do not cluster, prompting recalibration or root-cause analysis if they do. In this way, the EVT overlay stabilizes precisely the part of the limit structure where sampling noise is greatest, without diluting the empirical character of the central distribution used for day-to-day governance.

It is important to avoid static thresholds as they are operationally appealing but horizon-incoherent. Under a Normal proxy, the implied breach probability and quantile are:

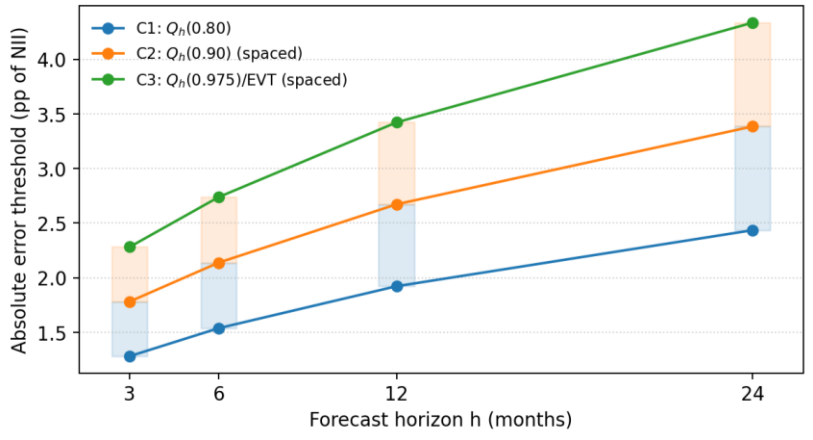
$$p_h\left(L^{static}\right) = \left(1 - \Phi\left(\frac{L^{static}}{\sigma_h}\right)\right), \quad \alpha_h\left(L^{static}\right) = 1 - p_h\left(L^{static}\right)$$

Here, L^{static} is the static one-sided tolerance band used as a horizon-invariant breach threshold for the forecast error. It's measured in the same units as the error, i.e., percentage points of NII. Because σ_h increases with h, the same L^{static} would map to different quantiles across horizons, which is undesirable characteristics for a limit system.

This statistical architecture - quantile-anchored tiers calibrated per horizon from the bank's empirical error distribution (with SEP-style RMSE mapping used only for communication), validated by unconditional and conditional coverage tests, with bias–dispersion separation to keep structural misspecification visible and a POT–GPD overlay to stabilize the deep tail - yields a single, horizon-aware framework that is both statistically coherent and governance-ready. The spacing discipline that separates adjacent tiers by a dispersion-scaled gap preserves materiality in committee actions and prevents artificial cascades, while periodic outcomes analysis under SR 11-7 ensures that realized exception frequencies and the independence of exceptions remain aligned with design targets.

Using Fed-consistent RMSE anchors interpolated to monthly horizons, we simulate forecast errors, compute the horizon-specific quantile tiers and the standardized exceedance score $S_{t,h} = | e_{t,h} | / Q_h(0.95)$, and benchmark EVT-based outer-tier limits against Normal-proxy limits at $h = 24$. The Figure 2 depicts how the calibrated tiers deliver coverage-consistent triggers at short horizons and why POT–GPD produces wider, more realistic deep-tail thresholds where long-horizon uncertainty is heaviest, thereby reducing false comfort from thin-tail assumptions and improving the reliability of the C3 boundary.

Figure 2. Horizon-specific tier limits with spacing (illustrative)



Because RMSE reflects the accumulation of multi-period forecast uncertainty, it typically increases in a convex pattern across horizons. In contrast, the calibrated escalation limits L_h derive from empirical quantiles and EVT-stabilized tails rather than from variance, and therefore grow more slowly than σ_h. When combined with the spacing discipline, the resulting tier curves are naturally smoother and less convex than the RMSE curve - a desirable property for horizon-aware governance.

3. Forecast-Uncertainty-Adjusted Breach Probabilities for ΔNII

A statistically coherent escalation regime begins by acknowledging that ΔNII are not point quantities but random variables whose realized values combine a structural projection with uncertainty arising from forecast error. Once this is recognized, limit monitoring becomes fundamentally probabilistic. The reported value at horizon h can be written as: $M_h = M_h^{(struct)} + \varepsilon_h$, where ε_h follows a horizon-specific empirical distribution $\mathcal{D}_h$ estimated from history or, for communication, a Normal proxy with variance σ_h^2 tied to the RMSE structure used in fan-chart practice. This formulation makes it explicit that the probability of a limit breach depends on the width of $\mathcal{D}_h$, which typically expands with the forecast horizon, rendering any fixed tolerance statistically inconsistent across h.

This perspective immediately transforms governance practice. Suppose we set a one-sided limit $L_h^{(k)}$ for the escalation level k. Under the RMSE-based Normal approximation, the breach probability and for a one-sided loss limit takes the form:

$$P_{breach}(h) = \Phi\left(\frac{-L_h^{(k)} - M_h^{(struct)}}{\sigma_h}\right)$$

Because σ_h increases with h, the same tolerance maps to very different quantiles across horizons. To restore coherence, the limit itself must vary with the horizon. If a bank wishes to impose the same exception probability $\bar{p}$ for all horizons, then the corresponding tolerance (one-sided) must satisfy:

$$T_h = Q_h(1 - \bar{p})$$

In short, ΔNII limits should be calibrated as horizon-specific quantiles derived from the institution's own forecast-error distribution, yielding escalation thresholds whose interpretation is stable across horizons and statistically tied to observed forecasting accuracy.

Within the risk-appetite construct, horizon-specific quantiles for ΔNII become the operational expression of tolerance and directly connect short-horizon planning to long-horizon steering. When a ΔNII outcome lies beyond the 90% or 97.5% quantile of its horizon-specific distribution, the breach can be interpretable as a low-frequency event under well-behaved forecasting performance. Because dispersion is embedded into the limit definition, breaches at long horizons are not implicitly discounted and breaches at short horizons are not inadvertently inflated. The resulting governance narrative replaces binary breaches with outcomes whose severity, recurrence, and independence can be read directly from the probability scale, aligning risk appetite with SR 11-7 expectations for outcomes analysis.

4. Conclusion and Forward Path

Several implications follow. First, comparability is restored across horizons because dispersion growth is embedded directly into the limit definition; short-horizon performance is not over-penalized and long-horizon outcomes are not under-weighted. Second, diagnostics become sharper: bias–dispersion decomposition separates structural misspecification from random variation, conditional-coverage testing detects clustering of exceptions, and EVT prevents brittleness in the outer tier. Third, governance and disclosure improve, since horizon-specific breach probabilities for ΔNII provide a transparent mapping between model uncertainty and risk appetite that is suitable for ALCO reporting and external communication.

The framework is not without limits. Long-horizon calibration is data-hungry and may require wider uncertainty bands, pooled-horizon estimation, or the use of informative priors during early adoption. Structural model changes can compromise the stationarity of the error process and therefore demand re-baselining and shadow backtests. Dependence structures for aggregation must be refreshed to avoid stale correlation and hidden tail dependence. Nevertheless, the empirical discipline of coverage testing and the modularity of the approach make these risks manageable, and, importantly, visible.

References

Basel Committee on Banking Supervision. 2016. Interest Rate Risk in the Banking Book. Basel: Bank for International Settlements. April.

Board of Governors of the Federal Reserve System, and Office of the Comptroller of the Currency. 2011. Supervisory Guidance on Model Risk Management (SR 11-7; OCC Bulletin 2011-12). Washington, DC.

Christoffersen, P. F. (1998). Evaluating interval forecasts. International Economic Review, 39(4), 841–862.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. Journal of Business & Economic Statistics, 13(3), 253–263.

Fissler, T., Frongillo, R., Hlavinová, J., & Rudloff, B. (2021). Forecast evaluation of quantiles, prediction intervals, and other set-valued functionals. Electronic Journal of Statistics, 15(1), 1034–1084.

Giacomini, R., & White, H. (2006). Tests of conditional predictive ability. Econometrica, 74(6), 1545–1578.

Pickands, J. (1975). Statistical inference using extreme order statistics. Annals of Statistics, 3(1), 119–131.

Reifschneider, D., & Tulip, P. (2017). Gauging the uncertainty of the economic outlook using historical forecasting errors: The Federal Reserve's approach. Finance and Economics Discussion Series 2017-020, Washington, D.C.: Board of Governors of the Federal Reserve System.