Evaluating Performance, Safety, and Robustness of an AI-Based Airport Delay Alerting Tool with Calibrated Machine Learning for Operational Decision Support

Special Session: PSRAI - Performance, Safety and Robustness in Artificial Intelligence-based Systems

Soufiane Momtaz, ENSET Mohammedia, Hassan II University of Casablanca, Morocco
Dr. Otmane Idrissi, ENSET Mohammedia, Hassan II University of Casablanca, Morocco
Dr. Joseph Machrouh, Thales LAS France, Rungis, France

Paper Number: 68009

IARIA

PESARO 2026
The Sixteenth International Conference on Performance, Safety and Robustness in Complex Systems and Applications
May 24, 2026 to May 28, 2026 – Venice, Italy

The Presenter : Soufiane Momtaz

Air Traffic Controller

+ 15 Years of Experience
+ Tower Control
+ Approcah / Radar Control

Air Traffic Management

+ Phd Candidate
+ Air Traffic Optimization
+ Machine Learning and Artificial Intelligence

The Lab: Statistical Success

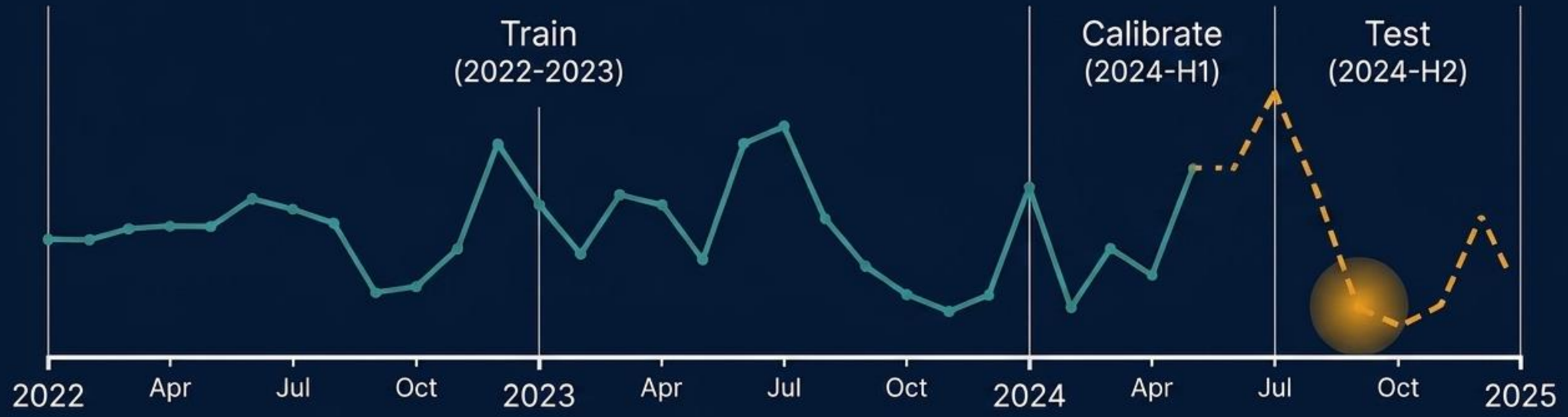
The model perfectly ranks historical data.
In validation, the AI looks flawless.

The Tarmac: Operational Risk

When operating conditions shift, that same highly accurate model triggers costly false alarms and silent failures.

High accuracy in the lab does not guarantee safe decisions on the tarmac. We must evaluate AI not as a statistical ranker, but as a cost-weighted operational decision engine.

The Danger of the Silent Shift



Airport environments change rapidly. A machine learning model calibrated in the first half of the year silently degrades as disruption prevalence shifts in the second half. The tool turns from an asset into a liability.

A New Paradigm for Operational AI

Performance

AUC, Brier Score

Can the model predict upcoming high-delay regimes?

Safety

Nominal-threshold risk

What is the operational cost of the alerting policy under a 5:1 (Missed Alert vs. False Alarm) cost ratio?

Robustness

Threshold-Local Calibration Error, Action-Overconfidence Gap

Does the system remain trustworthy when operating conditions or governance thresholds change?

True operational assurance lives exclusively in the center where all three overlap.

The Score-to-Action Pathway

The danger doesn't just live in the AI model. It lives in the space between the probability score and the governed action. Calibration is not a static fix; it is a sensitive operational control variable.

The Ultimate Operational Stress Test

The Scale

- 68,085 real airport-carrier-month records from the U.S. Bureau of Transportation Statistics (BTS).
- 377 U.S. Airports, 21 Airlines.

The Rules

- Zero Contemporaneous Leakage.
- Target-month outcomes are strictly excluded from the feature set.
- One-month-ahead forward predictions only.

The Threat Model

- Tested against Calibration-Transfer Failure.
- Evaluating when models trained on past data break down under shifting future prevalence.

Baseline Diagnostics: The Contenders

	Performance	Safety
Logistic Regression Good interpretability, trailing predictive quality.	AUC 0.763, Risk 0.612 	?
Random Forest Better predictive quality, higher threshold risk.	AUC 0.768, Risk 0.572 	?
XGBoost Highest AUC, Lowest Brier, Lowest Nominal-Threshold Risk.	AUC 0.811, Risk 0.515 	?

Uncalibrated, XGBoost wins across the board. But what happens when we try to make these models safer through standard post-hoc calibration?

THE CALIBRATION TRAP

VALIDATION (H1)

The illusion of safety.

Platt scaling brilliantly reduces Threshold-Local Calibration Error (TLCE) from 0.1018 to 0.0134.

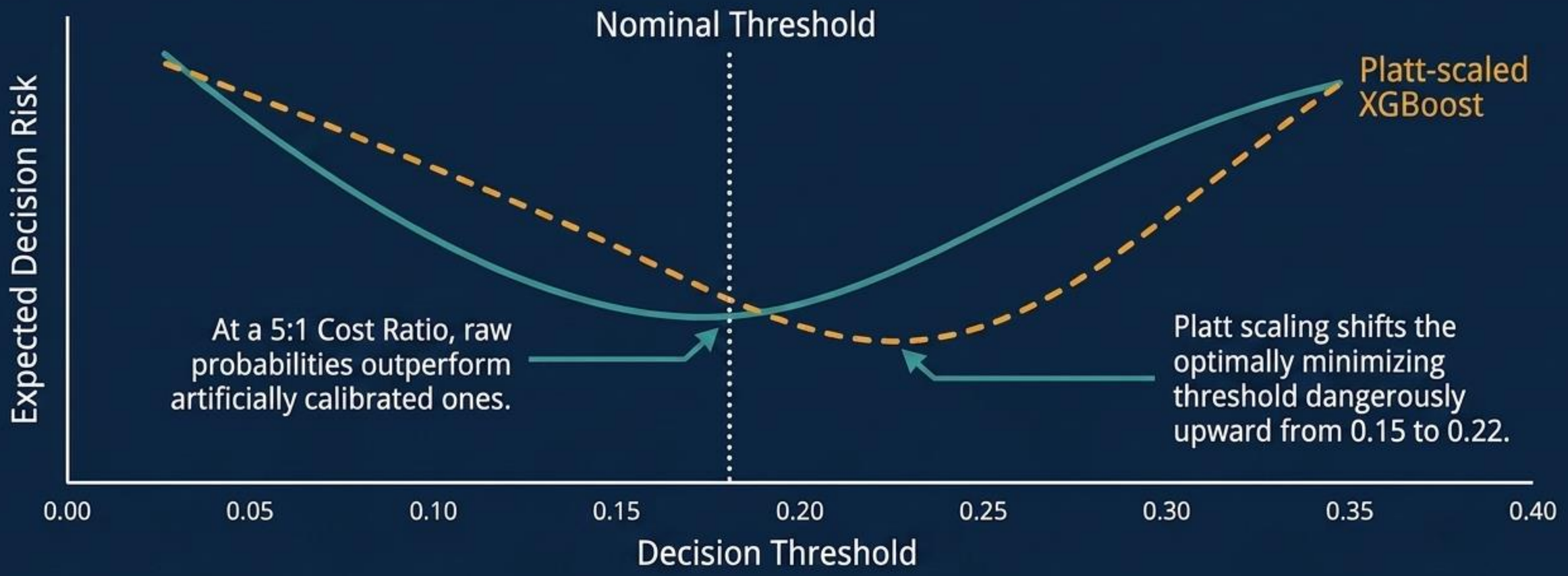
DEPLOYMENT (H2)

The reality of drift.

The exact same calibration map under **temporal** drift increases TLCE to 0.0441 and **worsens** decision risk from 0.5149 to 0.5266.

POST-HOC CALIBRATION ACTUALLY WORSENS SAFETY AND ROBUSTNESS WHEN THE OPERATING REGIME SHIFTS. IT IS NOT A SET-AND-FORGET FEATURE.

THE COST OF FALSE CONFIDENCE



Calibration silently changes the threshold that an operator finds most effective in practice, even when the nominal cost model is unchanged.

ANATOMY OF A TRANSFER FAILURE

THE SAFE ZONE (JULY 2024)

Condition: High delay prevalence (62.7%).

Result: Calibration helps.
Risk falls from 0.639 to 0.465.

THE DANGER ZONE (OCTOBER 2024)

Condition: Prevalence collapses (5.7%).

Result: Calibration severely harms safety. The calibrated model drastically over-triggers alerts, causing risk to spike from 0.285 to 0.427.

THE EFFECT OF CALIBRATION REVERSES ACROSS DEPLOYMENT SUB-REGIMES. WITHOUT THRESHOLD-LOCAL DIAGNOSTICS, THIS SILENT FAILURE GOES UNNOTICED.

THE PROFESSIONAL'S DEPLOYMENT CHECKLIST

WARNING SIGN	DIAGNOSIS	ACTION
Action rate rises while prevalence stays low.	System is over-alerting due to drift.	Review the alert threshold or freeze the calibration map until local fit is re-validated.
Calibration improves on fitting window but local error degrades on test.	Calibration-transfer failure.	Treat the calibrator as a time-sensitive component and re-audit on recent data.
Nominal-threshold risk and threshold-range risk disagree.	Policy-threshold mismatch.	Report both policy-specific safety and threshold-range robustness before deployment changes.

This research doesn't just offer an algorithm; it offers a secure, scalable blueprint for trustworthy AI in complex systems. It proves that calibration belongs inside the continuous safety loop, not outside it.

Evaluating Performance, Safety, and Robustness of an AI-Based Airport Delay Alerting Tool with Calibrated Machine Learning for Operational Decision Support

THANKS FOR ATTENTION

Contact the authors:

Soufiane Momtaz, ENSET Mohammedia, Hassan II University of Casablanca, Morocco
Email: soufiane.momtaz@gmail.com

Dr. Otmane Idrissi, ENSET Mohammedia, Hassan II University of Casablanca, Morocco
Email: iodrimane@gmail.com

Dr. Joseph Machrouh, Thales LAS France, Rungis, France
Email: joseph.machrouh@thalesgroup.com

Paper Number: 68009

PESARO 2026
The Sixteenth International Conference on Performance, Safety and Robustness in Complex Systems and Applications
May 24, 2026 to May 28, 2026 – Venice, Italy

Special Session: PSRAI - Performance, Safety and Robustness in Artificial Intelligence-based Systems