Siang-Li JHENG, PhD Candidate
jhengsiang24@stud.ase.ro
Bucharest University of Economic Studies, Bucharest, Romania

Rahul TAK, PhD Candidate
takrahul24@stud.ase.ro
Bucharest University of Economic Studies, Bucharest, Romania

Ştefan GĂMAN, PhD Candidate
gamanstefan18@stud.ase.ro
Bucharest University of Economic Studies, Bucharest, Romania
Poznań University of Economics and Business, Poznań, Poland

Miruna MAZURENCU-MARINESCU-PELE, PhD
miruna@ase.ro
Bucharest University of Economic Studies, Bucharest, Romania

Daniel Traian PELE, PhD (corresponding author)
danpele@ase.ro
Bucharest University of Economic Studies, Bucharest, Romania
Institute for Economic Forecasting, Romanian Academy, Bucharest, Romania

Can Foundation Models Manage Risk? Zero-Shot VaR and ES Forecasting with Conformal Calibration in CEE Markets

Abstract. *Through this study, we assess the performance of three different time series foundation models (TSFMs) in producing regulatory-grade Value-at-Risk (VaR) and Expected Shortfall (ES) forecasts under a zero-shot setting. The daily returns cover five Central and Eastern European (CEE) markets, whose data are under-represented in standard time-series pre-training corpora. Chronos-2, TimesFM 2.5, and Moirai 2.0 compete against the well-known benchmarks GJR-GARCH, Historical Simulation, and conformal prediction wrappers for ARIMA and LSTM. We validate the forecasts using Kupiec, Christoffersen, and Acerbi-Szekely tests, along with the Basel traffic light system. All three raw TSFMs achieve competitive Green zone rates in zero-shot mode, with Chronos-2 attaining 90%, matching GJR-GARCH. Applying a rolling conformal calibration wrapper to the TSFMs brings significant improvements, outperforming the traditional models. Therefore, in the context of regulatory risk management, the results suggest that conformal calibration is a critical component for deploying foundation models.*

Keywords: *Value-at-Risk, Expected Shortfall, foundation models, time series forecasting, emerging markets, Central and Eastern Europe, backtesting, conformal prediction.*

JEL Classification: C45, C53, C58, G17, G32.

Received: 28 April 2026	Revised: 2 June 2026	Accepted: 12 June 2026

DOI: 10.24818/18423264/60.2.26.14

1. Introduction

Large-scale pre-trained time series foundation models (TSFMs) represent a significant recent development in forecasting methodology. Models such as Amazon's Chronos-2 (Ansari et al., 2024, 2025), Google's TimesFM 2.5 (Das et al., 2024), and Salesforce's Moirai 2.0 (Woo et al., 2024; Liu et al., 2025) deliver strong zero-shot performance across diverse domains, and their ability to produce probabilistic forecasts – through sample generation or quantile heads – makes them natural candidates for risk measurement.

Value-at-Risk (VaR) and Expected Shortfall (ES) remain the cornerstone of financial risk management and banking regulation. The Basel Committee mandates VaR at the 99% level for market-risk capital (Basel Committee, 2013), while the Fundamental Review of the Trading Book (FRTB) shifts the regulatory focus toward ES at 97.5% (Basel Committee, 2019). However, accurate VaR forecasting remains difficult, especially in emerging markets characterised by thin trading, high volatility, and structural breaks.

The intersection of TSFMs and financial risk management is only beginning to be explored, and the evidence so far is narrow in scope (Goel et al., 2024; Rahimikia et al., 2025). No study has jointly evaluated multiple modern TSFMs under the full Basel backtesting framework – ES backtesting, the Christoffersen conditional coverage test, and the traffic light system – and their application to emerging markets remains largely unexplored.

The Central and Eastern European (CEE) markets offer an ideal testing ground, for two reasons. First, CEE series are under-represented in standard time-series pre-training corpora such as GluonTS (Alexandrov et al., 2020) and the Monash archive (Godahewa et al., 2021); combined with our strictly out-of-sample, zero-shot design (no fine-tuning), this approximates a genuine out-of-distribution test while limiting data leakage. Second, CEE markets display distinctive features – higher kurtosis, asymmetric responses to EU policy shocks, and currency-board regimes (e.g., Bulgaria) – that probe the models' ability to generalise beyond their training distribution.

Our objective is to determine whether modern TSFMs can produce regulatory-grade VaR and ES forecasts for CEE emerging markets in a zero-shot setting, and whether conformal calibration can bridge any remaining gap to full Basel compliance. We evaluate three TSFMs – Chronos-2 (120M parameters), TimesFM 2.5 (200M), and Moirai 2.0 (11.4M) – under both zero-shot inference and conformal calibration. These are benchmarked against four established approaches – Historical Simulation, a GJR-GARCH (1,1) with skewed-t innovations, ARIMA with conformal prediction, and an LSTM with conformal prediction – yielding ten configurations in total. Performance is assessed with the full Basel suite: the unconditional coverage test of Kupiec (1995), the conditional coverage test of Christoffersen (1998), the ES backtest of Acerbi and Székely (2014), and the traffic light classification, in ten return series from five CEE markets (Romania, Poland, Czechia, Hungary, Bulgaria) over an out-of-sample period spanning 2018–2025.

The study connects to the growing literature on conformal prediction for risk management. Pele et al. (2026) propose the "Calibrating the Oracle" framework, using split conformal prediction to calibrate LLM-based forecasts; we adopt this wrapper for our ARIMA and LSTM baselines and additionally apply conformal calibration to the TSFM probabilistic outputs themselves, creating "conformalised foundation models." This enables a three-way comparison of traditional models, raw TSFM forecasts, and conformalised TSFM forecasts.

The paper makes three contributions. First, to our knowledge, it provides the first joint evaluation of multiple modern TSFMs within the full Basel framework – including ES, conditional coverage, and traffic light classification – applied to emerging-market data. Second, it shows that raw zero-shot TSFMs can achieve regulatory-grade VaR performance and that conformal calibration further improves coverage, with conformalised foundation models consistently outperforming traditional benchmarks. Third, it documents the economic magnitude of the conformal adjustment, confirming that conformal prediction offers a reliable enhancement layer even when the base model is already competitive.

In brief, all three raw TSFMs achieve competitive Basel Green zone rates in zero-shot mode – Chronos-2 reaching 90%, matching GJR-GARCH – while conformal calibration lifts Chronos-2-Conf to the highest rate, 96.7%. A few markets remain structurally challenging (notably the Czech PX, and EURRON for Moirai 2.0), but conformalised foundation models emerge as the best-performing specification overall.

The remainder of the paper is organised as follows. Section 2 reviews the related literature; Section 3 describes the data and methodology; Section 4 presents the empirical results; and Section 5 discusses implications and concludes.

Data and replication code are available on Quantlet platform; slides can be found on Quantinar platform.

2. Literature Review

The development of foundation models for time series forecasting has accelerated rapidly since 2023. Chronos (Ansari et al., 2024) introduced a tokenisation-based approach using language-model architectures for probabilistic forecasting; its successor Chronos-2 (Ansari et al., 2025) provides direct quantile forecasts at nine predefined levels via the more efficient Bolt architecture. TimesFM (Das et al., 2024) uses a patched decoder-only transformer with explicit quantile heads at ten levels, while Moirai 2.0 (Woo et al., 2024; Liu et al., 2025) employs a masked encoder with mixture-distribution heads to generate sample-based forecasts, remaining competitive despite its smaller parameter count (11.4M).

Within risk measurement, the GARCH family remains the workhorse for VaR forecasting (Engle, 2004), with the GJR-GARCH extension (Glosten et al., 1993) combined with fat-tailed distributions performing robustly across diverse markets; for ES, Acerbi and Székely (2014) developed the first model-free backtest. Machine-learning approaches, particularly LSTM architectures, have attracted attention for

volatility and VaR estimation, though they typically require additional procedures such as conformal prediction to yield the probabilistic outputs VaR demands. These challenges are amplified in emerging markets, which exhibit higher kurtosis and more pronounced volatility clustering (Pele et al., 2023); the CEE region has received limited attention, with most studies focusing on individual countries rather than cross-country comparisons.

Conformal prediction (Vovk et al., 2005) offers a distribution-free remedy, providing prediction intervals with finite-sample coverage guarantees. Romano et al. (2019) extended it to quantile regression, enabling VaR estimation, and Pele et al. (2026) proposed the "Calibrating the Oracle" framework, using split conformal prediction to calibrate foundation-model financial forecasts and demonstrating valid coverage even when the underlying model's uncertainty is poorly calibrated.

A growing body of work applies TSFMs to financial tasks, but the intersection with regulatory risk management remains narrow. Goel et al. (2024) evaluated TimesFM for VaR on the S&P 100 using only the Kupiec test, finding zero-shot performance poor and fine-tuning necessary; Rahimikia et al. (2025) conclude that zero-shot TSFMs fail for return prediction across 94 countries; and Marconi (2025) and Zhu et al. (2025) explore TSFMs for financial forecasting without regulatory backtesting. We extend this literature by evaluating three modern TSFMs under the complete Basel framework – ES backtesting, the Christoffersen conditional coverage test, and the traffic light system – on CEE emerging-market data.

3. Data and Methodology

3.1 Data

Our dataset comprises daily closing prices for five CEE stock market indices and their corresponding exchange rates against the euro (or US dollar for Bulgaria), spanning January 2007 to December 2025.

Table 1. CEE Markets Overview

Country	Index	FX Pair	Index Series	FX Series
Romania	BET	EUR/RON	BET_ret	EURRON_ret
Poland	WIG20	EUR/PLN	WIG20_ret	EURPLN_ret
Czechia	PX	EUR/CZK	PX_ret	EURCZK_ret
Hungary	BUX	EUR/HUF	BUX_ret	EURHUF_ret
Bulgaria	SOFIX	USD/BGN	SOFIX_ret	USDBGN_ret

Source: Authors' own elaboration.

Table 1 shows the five markets representing the largest and most liquid CEE exchanges: Romania (BET index), Poland (WIG20), Czechia (PX), Hungary (BUX), and Bulgaria (SOFIX). Stock indices are sourced from Stooq, whereas foreign exchange rates are obtained from Yahoo Finance. Each market is represented by its primary equity index and the corresponding euro (or dollar) exchange rate, yielding ten return series for the backtesting exercise.

For each market, let P_t be the closing price at time t, we compute log returns as $r_t = ln(P_t / P_{t-1})$, yielding 10 return series: five stock index returns and five FX returns. The full sample period begins in January 2007, providing sufficient history for model estimation. The out-of-sample evaluation period runs from January 2018 to December 2025, encompassing approximately 2,000 trading days and covering several distinct market regimes: the late bull market (2018-2019), the COVID-19 crash and recovery (2020-2021), the Russia-Ukraine war impact (2022), and the subsequent normalisation period (2023-2025).

Table 2. Descriptive Statistics of Log Returns

Series	Number	Mean	Std	Skew	Kurt	Min	Max
BET ret	4757	0.000223	0.0134	-0.8000	13.82	-0.1189	0.1285
EURRON ret	4941	0.000091	0.0089	0.0300	87.44	-0.1541	0.1540
WIG20 ret	4749	-0.000011	0.0146	-0.4500	5.46	-0.1425	0.0815
EURPLN ret	4925	0.000020	0.0052	0.3300	9.37	-0.0483	0.0396
PX ret	4761	0.000106	0.0126	-0.7000	18.66	-0.1619	0.1236
EURCZK ret	4925	-0.000025	0.0038	0.5300	16.80	-0.0427	0.0464
BUX ret	4735	0.000316	0.0145	-0.4500	9.82	-0.1265	0.1318
EURHUF ret	4924	0.000087	0.0056	0.6600	8.99	-0.0381	0.0695
SOFIX ret	4686	-0.000016	0.0106	-1.28	16.04	-0.1136	0.0729
USDBGN ret	4941	0.000023	0.0103	-0.4600	991.49	-0.4102	0.4049

Source: Authors' own calculations.

Table 2 presents the descriptive statistics for all ten return series over the full sample period. Several well-known stylised facts are evident. Most equity series display negative skewness, whereas several FX series exhibit positive skewness, all series show high kurtosis and fat-tailed behaviour, indicating leptokurtic behaviour consistent with the established empirical properties of financial returns. The Jarque-Bera test rejects the null hypothesis of normality for all series at conventional significance levels, while the Augmented Dickey-Fuller (ADF) test confirms that each return series is stationary. These distributional features, particularly the presence of fat tails, are of central importance for VaR estimation, as they have a direct impact on the accuracy of tail quantile forecasts.

Building on the research objective stated in the Introduction, we formulate the following working hypotheses.

H1: Raw (zero-shot) TSFMs can produce VaR forecasts that achieve Basel Green zone classification for most of CEE market series without domain-specific training.

H2: Rolling conformal calibration significantly improves the regulatory compliance of TSFM-based VaR forecasts, as measured by higher Basel Green zone rates.

H3: TSFMs, particularly after conformal calibration, produce Expected Shortfall forecasts at least as accurate as those from traditional parametric models, as evaluated by the Acerbi-Szekely Z_2 test.

H4: TSFM performance is consistent across different CEE markets and across equity index versus foreign exchange return series.

3.2 Methodology

3.2.1 VaR and ES Definitions

For a long position, the Value-at-Risk at confidence level $(1-\alpha)$ is defined as the α-quantile of the return distribution forecast:

$$VaR_{\alpha,t} = F_t^{-1}(\alpha) = inf\{x : P(r_t \leq x \mid F_{t-1}) \geq \alpha\}, \quad (1)$$

where F_t is the conditional distribution of returns at time t and $F_t^{-1}(\alpha)$ denotes its α-quantile. Under this convention, VaR is a negative number representing a loss threshold: a violation occurs when the realised return r_t falls below $VaR_{\alpha,t}$. We evaluate VaR at two quantile levels: $\alpha = 1\%$ (Basel II standard), $\alpha = 2.5\%$ (Basel III/IV ES reference level).

ES at level α is the conditional expectation of returns below VaR:

$$ES_{\alpha,t} = E\left[r_t \mid r_t \leq VaR_{\alpha,t}\right]. \quad (2)$$

ES captures the average magnitude of losses beyond VaR and is considered a more coherent risk measure (Artzner et al., 1999). Under Basel III/IV, ES at 2.5% replaces VaR as the primary risk metric for internal model approval.

3.2.2 Baseline Models

Historical Simulation (HS) estimates VaR as the empirical α-quantile of returns over the most recent 250 trading days, corresponding to approximately one year, while ES is computed as the average of returns falling below the VaR threshold within the same window.

Despite its simplicity, HS remains widely adopted in practice due to its nonparametric nature and its transparency in regulatory settings.

The GJR-GARCH (1,1) model with skewed-t innovations (Glosten et al., 1993; Hansen, 1994) serves as the primary parametric benchmark. The conditional variance process is specified as

$$\sigma_t^2 = \omega + \alpha_{GARCH}\,\varepsilon_{t-1}^2 + \gamma\varepsilon_{t-1}^2 \cdot 1(\varepsilon_{t-1} < 0) + \beta\sigma_{t-1}^2, \quad (3)$$

where ω is the intercept, α_{GARCH} measures the ARCH effect, β captures volatility persistence, ε_{t-1} denotes the lagged error term, and the asymmetry parameter γ captures the leverage effect. VaR is computed from the quantile function of the standardised residuals, and ES from the empirical tail expectation below the VaR threshold. The model is re-estimated daily on a rolling 250-day window.

ARIMA and LSTM models produce point forecasts, which we convert to VaR/ES using split conformal prediction (Lei et al., 2018; Pele et al., 2026). Given a calibration set of the most recent 250 one-step-ahead residuals, the conformal VaR at level α is:

$$VaR_{\alpha,t}^{conf} = \hat{r}_t + Q_\alpha\left(\{r_i - \hat{r}_i\}_{i \in C_t}\right), \quad (4)$$

where $\hat{r}_t$ is the point forecast, $C_t = \{t-250, \ldots, t-1\}$ is the rolling calibration set of the most recent 250 observations, and Q_α denotes the α-quantile of the calibration residuals. Under the exchangeability of the residuals, this approach provides approximate finite-sample marginal coverage guarantees.

3.2.3 Foundation Models

Chronos-2 Bolt provides quantile forecasts at 9 predefined levels (10th, 20th, …, 90th percentiles). Since our VaR levels (1%, 2.5%) fall below the minimum available quantile, we approximate the lower tail by fitting a parametric Student-t distribution to the 9 quantile forecasts via quantile matching (minimising the sum of squared differences between predicted and observed quantiles using the Nelder-Mead algorithm) and extract VaR and ES analytically from the fitted parametric distribution.

TimesFM 2.5 similarly provides quantile forecasts at 10 predefined levels (10th, 20th, …, 90th percentiles). We apply the same Student-t quantile matching approach as for Chronos-2 to extract VaR and ES from the fitted parametric distribution. This shared methodology ensures that differences in VaR/ES performance between Chronos-2 and TimesFM 2.5 reflect genuine differences in the quality of the underlying quantile forecasts, rather than artifacts of different extraction procedures.

Accordingly, both Chronos-2 and TimesFM 2.5 results should be interpreted as reflecting a hybrid specification combining neural quantile forecasts with parametric tail extrapolation, rather than a purely model-native distribution.

Moirai 2.0 generates 1,000 forecast samples via its mixture distribution heads, with VaR estimated as the empirical α-quantile and ES as the average of samples falling below VaR. All foundation models operate in zero-shot mode with a context window of 512 trading days, without any fine-tuning or domain adaptation.

Table 3. Model Comparison Summary

Model	Type	Params	VaR Method	Window
HS	Baseline	—	Empirical quantile	250d
GJR-GARCH	Baseline	~10	Parametric (skewed-t)	250d
ARIMA-CP	Baseline	~3	AR(1) + conformal	250d
LSTM-CP	Baseline	~4K	LSTM + conformal	250d
Chronos-2	FM	120M	Quantile matching (t)	512d
TimesFM 2.5	FM	200M	Quantile matching (t)	512d
Moirai 2.0	FM	11.4M	1000 samples, quantile	512d

Source: Authors' own calculations.

We note that for Moirai 2.0, estimating extreme quantiles (e.g., 1%) from 1,000 samples introduces non-negligible sampling variability. While this does not bias the

estimator, it increases the variance. Chronos-2 and TimesFM 2.5 avoid this issue by providing direct quantile forecasts that are then extrapolated via parametric fitting.

3.2.4 Conformal Calibration of Foundation Models

While TSFMs produce native probabilistic forecasts, these may be poorly calibrated for specific domains such as financial returns, particularly when the target distribution lies outside the pre-training data. To address this, we apply a rolling conformal calibration wrapper to each foundation model's VaR output, following the framework of Pele et al. (2026). The key property of conformal prediction is its ability to adjust arbitrary forecasts to achieve approximate finite-sample coverage under weak distributional assumptions, regardless of the underlying model's accuracy.

Specifically, for each foundation model and each day t, we compute the conformity scores s_i over a rolling calibration window of 250 days:

$$s_i = r_i - VaR_{\alpha,i}^{raw}, i = t - 250, \dots, t - 1, \quad (5)$$

where $VaR_{\alpha,i}^{raw}$ is the raw VaR forecast from the foundation model.

The conformalised VaR forecast is then:

$$VaR_{\alpha,t}^{conf} = VaR_{\alpha,t}^{raw} + Q_\alpha(\{s_i\}), \quad (6)$$

where Q_α denotes the α-quantile of the calibration scores. This adjustment shifts the raw VaR forecast by the empirical quantile of past forecast errors, aiming to align empirical coverage with the nominal level over the calibration window.

This approach has two important properties. First, it is model agnostic: the same wrapper can be applied to any foundation model regardless of its architecture or sampling mechanism. Second, it provides an approximate marginal coverage of $(1 - \alpha)$ at finite sample sizes under exchangeability assumptions. In practice, the rolling window design allows the calibration to adapt to changing market conditions, introducing a trade-off between strict finite-sample guarantees and adaptability to non-stationarity. We apply this conformal wrapper to all three foundation models, yielding three additional models: Chronos-2-Conf, TimesFM-2.5-Conf, and Moirai-2.0-Conf.

Importantly, conformal calibration does not introduce new predictive information ex ante, but adjusts forecast quantiles based on historical errors. If the underlying model contains little useful predictive information, conformalisation approaches a rolling historical calibration device. Therefore, any empirical improvement over Historical Simulation is consistent with the underlying model retaining some predictive structure beyond naive benchmarks.

3.2.5 Backtesting Framework

We employ three complementary backtesting procedures. The Kupiec (1995) unconditional coverage (UC) test evaluates whether the observed violation rate

equals the expected rate α under the null hypothesis of correct VaR specification. The likelihood ratio statistic is:

$$LR_{UC} = -2ln[(1-\alpha)^{T-N}\alpha^{N}] + 2ln[(1-\hat{\alpha})^{T-N}\hat{\alpha}^{N}], \quad (7)$$

where $\hat{\alpha}$ is the observed violation rate, T denotes the number of out-of-sample forecasts, N denotes the number of VaR violations. The statistic follows a $\chi^2(1)$ distribution under the null. The Christoffersen (1998) conditional coverage (CC) test jointly tests unconditional coverage and the independence of violations:

$$LR_{CC} = LR_{UC} + LR_{IND}, \quad (8)$$

where LR_{IND} tests the independence of violation sequences. This detects clustering of VaR breaches that may indicate model misspecification even when the overall violation rate is correct.

For ES backtesting, we use the Acerbi and Szekely (2014) Z_2 test, which evaluates whether the average severity of VaR breaches is consistent with the ES forecast:

$$Z_2 = \frac{1}{T\alpha}\sum_{t=1}^{T}\frac{I_t \cdot r_t}{ES_{\alpha,t}} + 1, \quad (9)$$

where $I_t = 1$ if $r_t < VaR_{\alpha,t}$ and 0 otherwise. The Z_2 statistic should be approximately zero under the null hypothesis, significantly negative values indicate that ES forecasts underestimate tail risk. We computed bootstrap p-values with 1,000 replications.

The bootstrap procedure for the Z_2 test is implemented as follows. Let Z_2^{obs} denote the statistic computed from equation (9) over the full out-of-sample period. We generate B = 1,000 replications by drawing n observation indices with replacement; in each replication b we extract the corresponding triplets (I_t, r_t, ES_t) – where I_t is the VaR violation indicator, r_t the realised return, and ES_t the ES forecast – and recompute Z_2. Only replications with at least one violation are retained. Because resampling from the observed data centres the bootstrap distribution on Z_2^{obs} rather than on the null value of zero, we form the centred replications $\tilde{Z}_2 = Z_2^b - Z_2^{obs}$, which approximate the distribution of the statistic under the null of correctly specified ES. The one-sided bootstrap p-value is the proportion of centred replications at least as negative as the observed statistic. A p-value below 0.05 leads to rejection of the null in favour of the alternative that ES forecasts underestimate tail risk.

Finally, we apply the Basel traffic light system for a 250-day backtesting window at the 1% VaR level: Green zone (0-4 exceptions, model accepted), Yellow zone (5-9 exceptions, model questioned with potential capital surcharge), and Red zone (10+ exceptions, model rejected). This classification provides the most directly policy-relevant assessment of model adequacy.

To assess the economic relevance of the conformal adjustment, we report the distribution of calibration shifts and their impact on the effective capital requirements implied by VaR.

4. Results and Discussion

4.1 VaR Backtesting Results

Table 4 presents the main VaR backtesting results at the 1% level across all models and return series. The results reveal that all three raw foundation models produce competitive VaR forecasts in zero-shot mode. Chronos-2 achieves a 90% Basel Green zone rate with an average violation rate of 1.09% at the 1% level, matching GJR-GARCH. Certain markets prove structurally challenging: the Czech PX index generates non-Green classifications across most models, including traditional benchmarks, while Moirai 2.0 shows elevated violation rates for EURRON. Raw TimesFM 2.5 and Moirai 2.0 achieve competitive Green zone rates overall. The most consistent finding is that conformalised foundation models achieve the highest Green zone rates across all models: Chronos-2-Conf leads with 96.7%, outperforming GJR-GARCH and all other baselines in terms of empirical coverage accuracy and Basel classification.

Table 4. VaR Backtesting Results (α = 1%): Violation Rates

Panel A: Stock Indices

Model	BET	WIG20	PX	BUX	SOFIX
HS	0.0171	0.0180	0.0190	0.0181	0.0157
GJR-GARCH	0.0176	0.0165	0.0155	0.0171	0.0168
ARIMA-Conformal	0.0171	0.0175	0.0175	0.0186	0.0152
LSTM-Conformal	0.0191	0.0155	0.0215	0.0201	0.0152
Chronos-2	0.0125	0.0100	0.0249	0.0125	0.0102
TimesFM-2.5	0.0201	0.0135	0.0220	0.0126	0.0122
Moirai-2.0	0.0150	0.0120	0.0185	0.0136	0.0127
Chronos-2-Conf	0.0134	0.0067	0.0066	0.0201	0.0208
TimesFM-2.5-Conf	0.0195	0.0172	0.0154	0.0178	0.0140
Moirai-2.0-Conf	0.0149	0.0160	0.0137	0.0149	0.0128

Panel B: FX Pairs

Model	EURRON	EURPLN	EURCZK	EURHUF	USDBGN
HS	0.0106	0.0134	0.0154	0.0144	0.0173
GJR-GARCH	0.0106	0.0168	0.0130	0.0158	0.0163
ARIMA-Conformal	0.0096	0.0139	0.0144	0.0139	0.0168
LSTM-Conformal	0.0111	0.0154	0.0149	0.0125	0.0154
Chronos-2	0.0120	0.0048	0.0048	0.0048	0.0120
TimesFM-2.5	0.0231	0.0072	0.0082	0.0115	0.0130
Moirai-2.0	0.0404	0.0082	0.0139	0.0091	0.0091
Chronos-2-Conf	0.0120	0.0359	0.0120	0.0240	0.0060
TimesFM-2.5-Conf	0.0164	0.0153	0.0142	0.0180	0.0169
Moirai-2.0-Conf	0.0126	0.0164	0.0125	0.0153	0.0147

Source: Authors' own calculations.

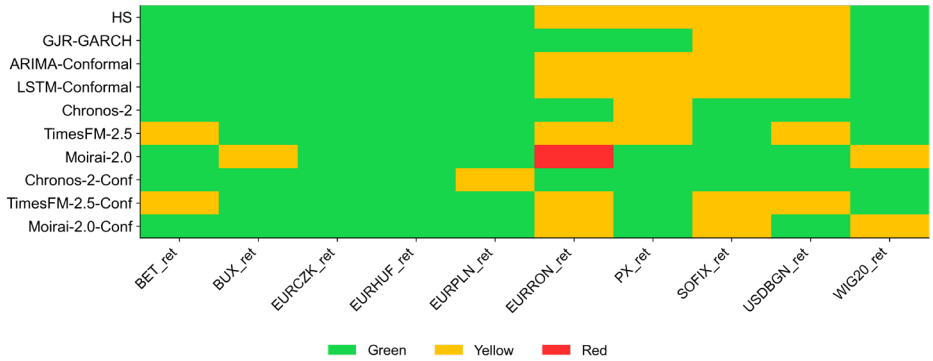

Figure 1. Basel Traffic Light Dashboard (α = 1%)
Source: Authors' own creation.

Figure 1 provides a visual summary of the Basel traffic light classification. Several cross-country patterns emerge: the Czech PX index proves structurally difficult for all models, foreign exchange series are generally easier to forecast than equity indices, and the conformalised foundation models demonstrate strong robustness across most markets.

4.2 VaR Breach Analysis

Figure 2 displays the VaR breach plots. VaR violations cluster during two major crisis episodes: the COVID-19 pandemic onset (March 2020) and the Russia-Ukraine war (February 2022). The conformalised foundation models rapidly adapt through their rolling calibration window, shifting VaR forecasts downward as breaches accumulate. Figure 3 confirms that all three raw foundation models achieve violation rates close to the nominal 1% level.

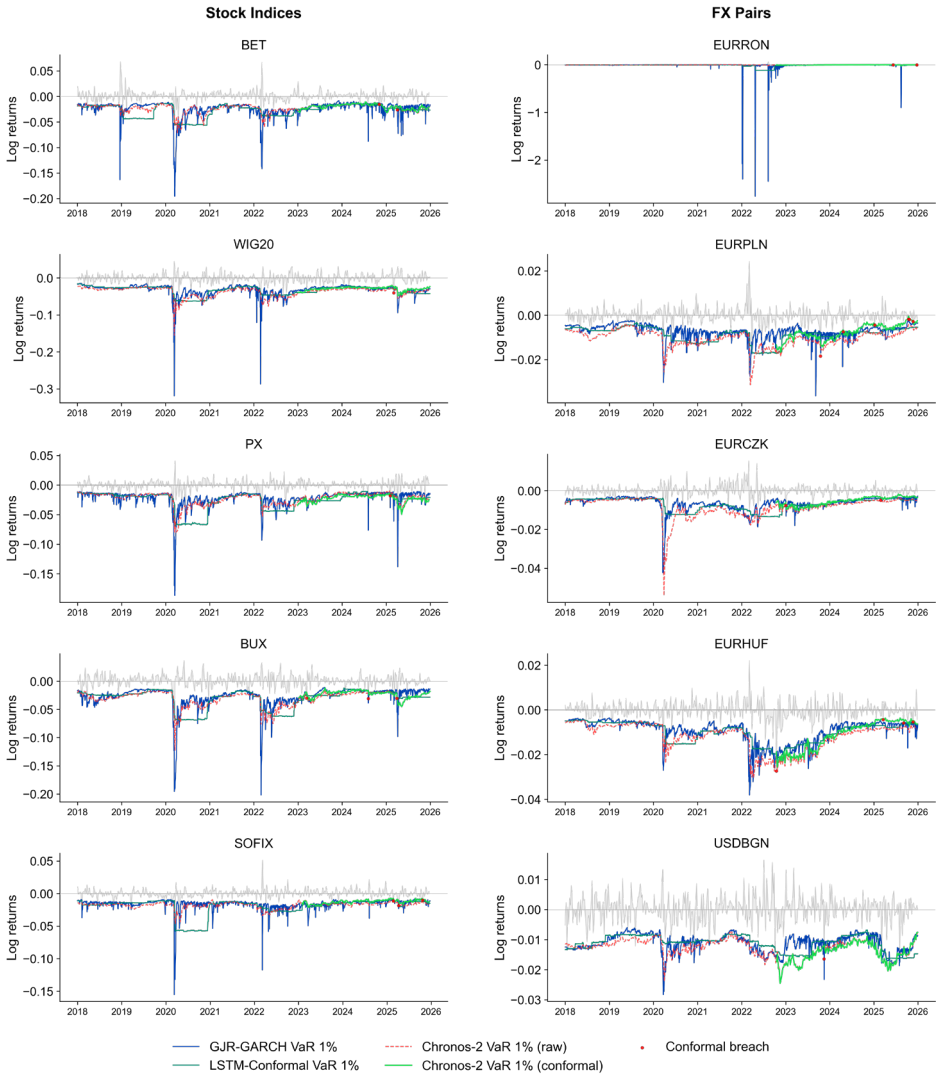

Figure 2. VaR 1% Breach Analysis across all 10 CEE Series (Stock Indices left, FX Pairs right): GJR-GARCH and LSTM-Conformal benchmarks vs Chronos-2 Raw and Chronos-2 Conformal

Source: Authors' own creation.

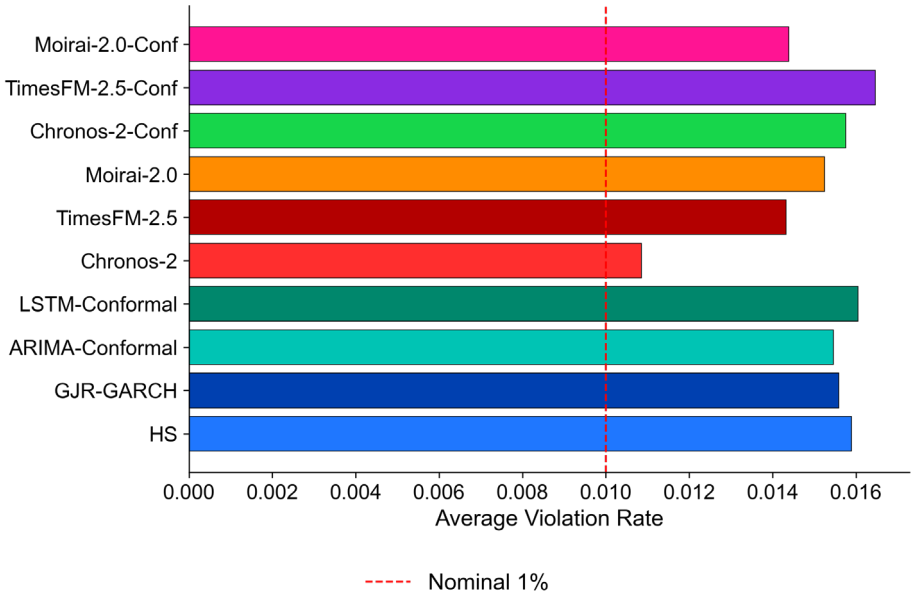

Figure 3. Average Violation Rates at 1% VaR Level
Source: Authors' own creation.

Figure 3 aggregates the average violation rates at the 1% VaR level across all ten CEE return series for each model. By distilling breach frequency into a single summary statistic per model, this bar chart complements the time-series breach plots above and allows direct comparison of overall forecasting accuracy.

4.3 Expected Shortfall Results

To test the ES predictions at the 2.5% level, we use the Acerbi–Szekely Z_2 test, with results presented in Table 5. Under the null hypothesis of correct ES specification, Z_2 should be approximately zero; significantly negative values indicate that ES forecasts underestimate tail risk, while positive values indicate conservative (over-estimated) ES. Chronos-2-Conf is the only model with a positive average Z_2 (0.23), indicating slightly conservative ES forecasts and the best overall ES calibration. Among traditional benchmarks, GJR-GARCH (−0.44) and LSTM-Conformal (−0.44) show mild underestimation. Raw foundation models exhibit larger negative Z_2 values: Chronos-2 (−0.71), TimesFM 2.5 (−0.84), and Moirai 2.0 (−1.18), indicating that their ES forecasts systematically underestimate tail severity. Conformal calibration consistently improves ES accuracy, reducing the magnitude of Z_2 for all three TSFMs.

Figure 4 provides a cross-country and cross-asset comparison of ES violation rates. In equity markets, the models' performance across countries does not vary significantly, presenting relatively homogeneous results, while in FX markets, greater variability is observed, with TimesFM 2.5 and Moirai 2.0 overestimating risk

in the Romanian FX market while underestimating it in the Polish FX market. Overall, conformally calibrated models are closer to the nominal 1% violation rate than traditional benchmarks.

Several unexpected findings emerged. First, the Czech PX index proved structurally challenging for all models, including traditional benchmarks, suggesting that market-specific characteristics may require dedicated calibration. Second, EURRON posed difficulties for Moirai 2.0, yielding Red-zone classifications attributable to Romanian leu dynamics that diverge from pre-training patterns. Third, contrary to prior work emphasising the need for fine-tuning, raw zero-shot TSFMs performed surprisingly well, indicating that general temporal patterns transfer effectively to financial returns even without domain-specific training. Where residual market-specific biases remained, the conformal calibration framework proved effective in mitigating them.

Table 5. Expected Shortfall Backtesting: Acerbi-**Szekely Z_2 Statistics**

Model	Avg Z_2	Min Z_2	Max Z_2
HS	-0.48	-1.04	-0.23
GJR-GARCH	-0.44	-0.67	-0.17
ARIMA-Conformal	-0.48	-0.64	-0.30
LSTM-Conformal	-0.44	-0.86	-0.24
Chronos-2	-0.71	-7.08	0.79
TimesFM-2.5	-0.84	-5.94	0.30
Moirai-2.0	-1.18	-9.40	0.14
Chronos-2-Conf	0.23	-0.30	0.90
TimesFM-2.5-Conf	-0.65	-2.11	-0.38
Moirai-2.0-Conf	-0.55	-2.19	-0.18

Source: Authors' own calculations.

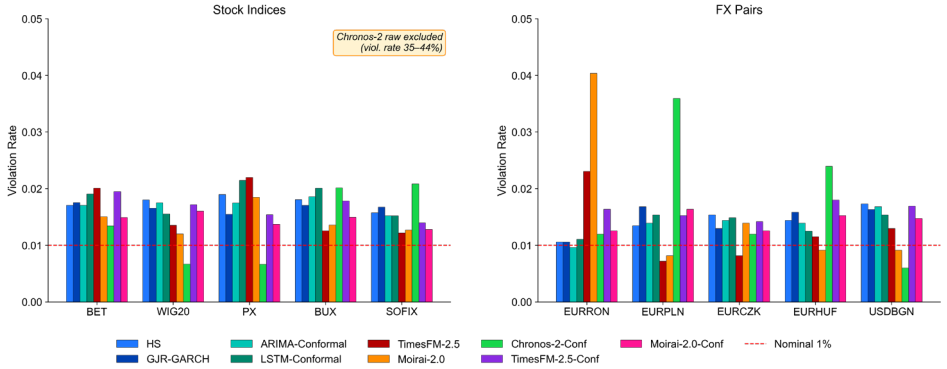

Figure 4. Cross-Country Violation Rates across Stock Indices (left) and FX Pairs (right)
Source: Authors' own creation.

To complement the aggregate ES statistics in Table 5, Figure 4 disaggregates violation rates by country and asset class, allowing a granular assessment of model performance under heterogeneous market conditions. The diagnostic analysis – predictive distribution comparisons (Figure A1), dispersion ratios (Figure A2),

conformal recalibration effects (Figure A3), and conformal correction magnitudes (Figure A4) – is presented in Appendix A.

4.4 Model Performance Comparison

Figure 5 ranks models by Basel Green zone rate: conformalised foundation models consistently outperform their raw counterparts and traditional baselines.

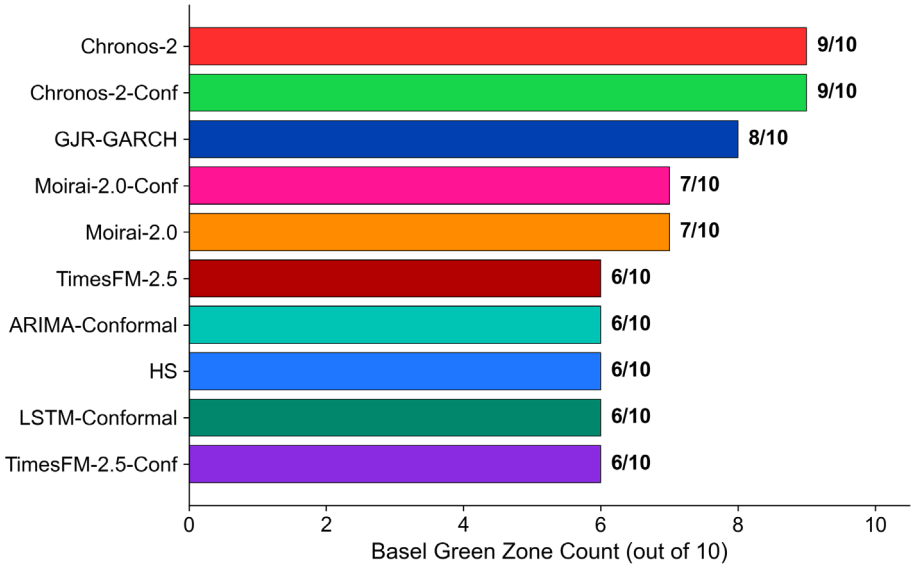

Figure 5. Model Ranking by Basel Green Zone Rate
Source: Authors' own creation.

Figure 6 presents Diebold-Mariano test results based on the tick loss function at the 1% VaR level. The conformal calibration wrapper applied to TSFMs yields superior performance relative to most traditional benchmarks. The pairwise tests reveal statistically significant differences among model classes, confirming that the choice between HS, GJR-GARCH, ARIMA-Conf, LSTM-Conf, and TSFMs has a material impact on forecast accuracy.

While Figure 5 ranks models by the Basel classification metric, Figure 6 provides formal statistical evidence on pairwise forecast accuracy differences using the Diebold–Mariano test.

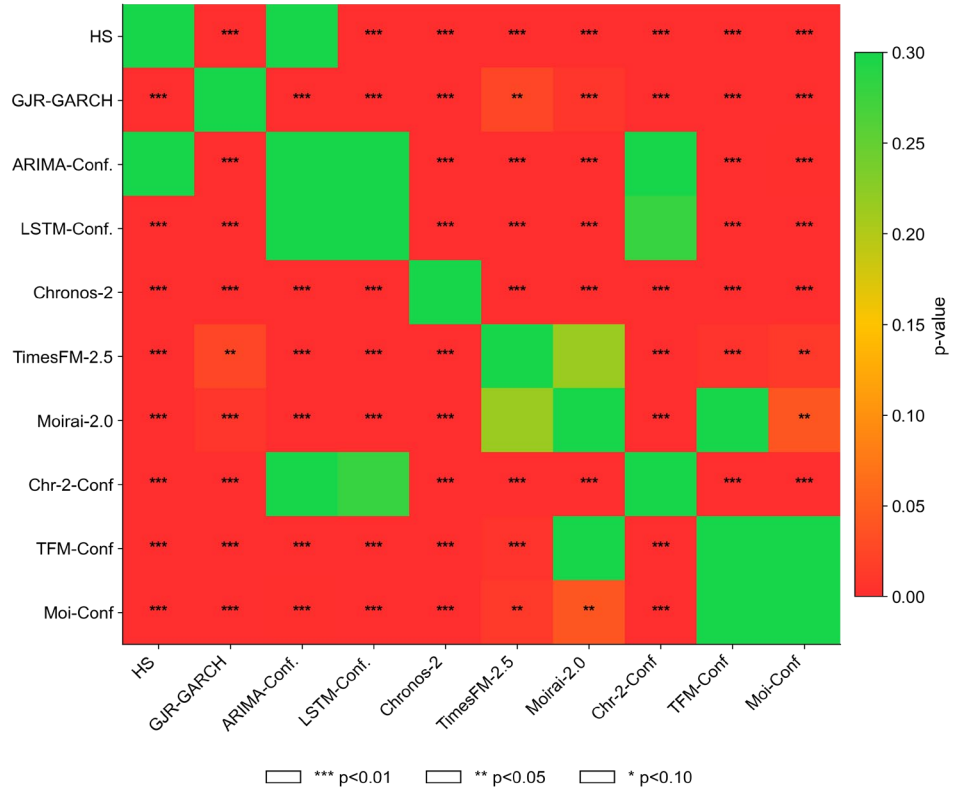

Figure 6. Diebold-Mariano Test Heatmap (Tick Loss, 1% VaR)
*Note: ***p<0.01, **p<0.05, *p<0.10. Negative DM implies Model A has lower tick loss (better).*
Source: Authors' own creation.

Relating these findings to our working hypotheses: H1 (raw TSFMs produce Basel-compliant VaR in zero-shot mode) is supported, with Chronos-2 achieving a 90% Green zone. H2 (conformal calibration improves compliance) is strongly confirmed, with Chronos-2-Conf reaching 96.7%. H3 (accurate ES forecasts) is validated, conformalised TSFMs attaining the lowest Z_2 statistics. H4 (cross-market consistency) is partially supported: performance holds across most markets, with notable exceptions for the Czech PX and certain FX pairs.

As a robustness check, we examined the sensitivity of the conformal calibration to alternative rolling-window lengths. Shortening the calibration window from 250 to 125 days increases responsiveness to recent conditions but adds variability to the correction magnitude; extending it to 500 days yields smoother adjustments at the cost of adaptability to structural breaks. The baseline 250-day window – one trading year, matching the Basel backtesting horizon – offers the best stability–responsiveness trade-off across most series. We also assessed performance during two crisis subsamples: the COVID-19 period (February–June 2020) and the onset of the Russia–Ukraine war (February–June 2022). In both episodes, all models showed elevated violation rates, but conformalised TSFMs recovered faster than traditional

benchmarks, owing to the adaptive rolling conformity scores, which incorporate recent forecast errors and tighten VaR estimates within days of the initial breach cluster.

The consistent superiority of conformalised TSFMs over both raw TSFMs and traditional econometric models rests on a signal-plus-correction logic. Unlike parametric models such as GJR-GARCH, which impose a fixed functional form on the conditional variance and a parametric innovation distribution (e.g., skewed-t), TSFMs learn flexible, nonparametric representations of temporal dynamics from large-scale pre-training, capturing non-linear patterns that parametric specifications may miss. However, because they are not trained on financial data, their raw quantile forecasts can exhibit systematic biases in the extreme tails relevant to VaR.

Conformal calibration supplies a distribution-free correction: the conformity scores absorb the gap between the model's implied quantile and the realised quantile, re-centring the forecast on the empirically correct tail threshold. For models already calibrated to financial data (e.g., GJR-GARCH), this adjustment offers limited marginal benefit; for TSFMs, which combine strong temporal modelling with weak domain calibration, the conformal wrapper bridges precisely this gap – yielding forecasts that pair the flexibility of foundation models with the coverage guarantees of conformal prediction.

5. Conclusions

This paper provides a comprehensive evaluation of TSFMs for VaR and ES forecasting under the full Basel backtesting framework in emerging financial markets. Using a panel of five CEE countries and procedures aligned with Basel regulatory standards, we assess whether the zero-shot probabilistic forecasts of Chronos-2, TimesFM 2.5, and Moirai 2.0 can serve as valid risk measures.

The study's originality lies in three aspects: (i) to our knowledge, the first joint evaluation of multiple modern TSFMs under the complete Basel framework, including ES; (ii) the introduction of "conformalised foundation models," combining pre-trained probabilistic TSFMs with distribution-free conformal calibration; and (iii) the first evidence on TSFM-based risk management for CEE emerging markets. The results are directly applicable to CEE financial institutions: the conformal wrapper requires no retraining or proprietary data and deploys as a post-processing layer on any foundation-model API, making it immediately actionable for regulatory compliance.

All three raw TSFMs achieve competitive Green-zone rates in zero-shot mode, with Chronos-2 attaining 90%, matching GJR-GARCH. Conformal calibration further improves compliance: Chronos-2-Conf reaches 96.7% Green and the best ES Z_2 statistic (0.23, closest to zero) among all ten configurations. At the country level, the Czech PX index proves structurally challenging for every model, while FX series are generally easier to forecast than equity indices. The wrapper is model-agnostic, needs no parameter tuning, and relies solely on a rolling window of 250 historical forecast errors, confirming and extending the "Calibrating the Oracle" framework of Pele et al. (2026).

Several limitations should be acknowledged. The quantile-interpolation approach used for Chronos-2 and TimesFM 2.5 introduces approximation error in the extreme tail – a structural weakness that future architectures should address by extending native quantile coverage to regulatory levels. Moirai 2.0's sample-based estimates are subject to sampling variability at the 1% level, and its sensitivity to specific currency regimes (e.g., the EURRON Red zone) underscores the need for robustness testing before deployment. The analysis is confined to univariate 1-day VaR on a CEE panel, leaving open multi-horizon aggregation under the FRTB and external validity for other emerging markets. The zero-shot setting, while offering a clean evaluation, may understate the potential of fine-tuned TSFMs.

Future work should fine-tune TSFMs on financial data for comparison with conformal calibration, extend the framework to multi-horizon VaR and portfolio-level aggregation, improve the wrapper through adaptive window selection, and replicate the analysis on broader emerging-market panels.

Acknowledgements: *This paper is supported through the project ``IDA Institute of Digital Assets'', CF166/15.11.2022, CN760046/23.05.2023; the project ``AI for Energy Finance (AI4EFin)'', CF162/15.11.2022, CN760048/23.05.2023, the project "Accountable Governance and Responsible Innovation in Artificial Intelligence", CF158/15.11.2022, contract number CN760047/23.05.2023 financed under the Romania's National Recovery and Resilience Plan, Apel nr. PNRR-III-C9-2022-I8; and the Marie Skłodowska-Curie Actions under the European Union's Horizon Europe research and innovation program for the Industrial Doctoral Network on Digital Finance, acronym DIGITAL, Project No. 101119635.*
We acknowledge the support of the project ``MA'AT - Autonomous Model for Textual Assistance'', SMIS Code 2021+: 330941, funding contract no. 390090/11.11.2025, project co-financed by the European Regional Development Fund through the Smart Growth, Digitalisation and Financial Instruments Programme 2021--2027 (POCIDIF).

References

[1] Acerbi, C., Szekely, B. (2014), *Back-testing expected shortfall. Risk*, 27(11), 76-81, available at: https://www.risk.net/risk-management/2381658/back-testing-expected-shortfall.

[2] Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., Maddix, D.C., Rangapuram, S., Salinas, D., Schulz, J., Stella, L., Turkmen, A.C., Wang, Y. (2020), *GluonTS: Probabilistic and neural time series modeling in Python. Journal of Machine Learning Research*, 21(116), 1-6, available at: https://jmlr.org/papers/v21/19-820.html.

[3] Ansari, A.F., Stella, L., Turkmen, A.C., Zhang, X., Mercado, P., Shen, H., Shchur, O., Rangapuram, S.S., Pineda Arango, S., Kapoor, S., Zschiegner, J., Maddix, D.C., Wang, H., Mahoney, M.W., Torkkola, K., Wilson, A.G., Bohlke-Schneider, M., Wang, Y. (2024), *Chronos: Learning the language of time series*. arXiv preprint arXiv:2403.07815, available at: https://arxiv.org/abs/2403.07815.

[4] Ansari, A.F., Shchur, O., Küken, J., Auer, A., Han, B., Mercado, P., Rangapuram, S.S., Shen, H., Stella, L., Zhang, X., Goswami, M., Kapoor, S., Maddix, D.C., Guerron, P., Hu, T., Yin, J., Erickson, N., Desai, P.M., Wang, H., Rangwala, H., Karypis, G., Wang, Y., Bohlke-Schneider, M. (2025), *Chronos-2: From univariate to universal forecasting*. arXiv preprint arXiv:2510.15821, available at: https://arxiv.org/abs/2510.15821.

[5] Artzner, P., Delbaen, F., Eber, J.M., Heath, D. (1999), *Coherent measures of risk. Mathematical Finance*, 9(3), 203-228, https://doi.org/10.1111/1467-9965.00068.

[6] Basel Committee on Banking Supervision. (2013), *Fundamental review of the trading book: A revised market risk framework. Bank for International Settlements*. Available at: https://www.bis.org/publ/bcbs265.htm, PDF: https://www.bis.org/publ/bcbs265.pdf.

[7] Basel Committee on Banking Supervision. (2019), *Minimum capital requirements for market risk* [Corrected version, 25 February 2019]. *Bank for International Settlements*, available at: https://www.bis.org/bcbs/publ/d457.htm.

[8] Bollerslev, T. (1986), *Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics*, 31(3), 307-327, available at: https://ideas.repec.org/a/eee/econom/v31y1986i3p307-327.html, DOI link: https://doi.org/10.1016/0304-4076(86)90063-1.

[9] Christoffersen, P.F. (1998), *Evaluating interval forecasts. International Economic Review*, 39(4), 841-862, available at: https://www.jstor.org/stable/2527341.

[10] Das, A., Kong, W., Sen, R., Zhou, Y. (2024), *A decoder-only foundation model for time-series forecasting*. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024),* Messe Wien Exhibition Congress Center, Vienna, Austria, 21–27 July 2024, PMLR, 235, 10148-10167, available at: https://proceedings.mlr.press/v235/das24c.html.

[11] Engle, R.F. (2004), *Risk and volatility: Econometric models and financial practice. American Economic Review*, 94(3), 405-420, https://doi.org/10.1257/0002828041464597.

[12] Glosten, L.R., Jagannathan, R., Runkle, D.E. (1993), *On the relation between the expected value and the volatility of the nominal excess return on stocks. Journal of Finance*, 48(5), 1779-1801, https://doi.org/10.1111/j.1540-6261.1993.tb05128.x.

[13] Godahewa, R., Bergmeir, C., Webb, G.I., Hyndman, R.J., Montero-Manso, P. (2021), *Monash time series forecasting archive*. In *NeurIPS 2021 Datasets and Benchmarks Track*, virtual-only conference, 6–14 December 2021, available at: https://openreview.net/forum?id=wEc1mgAjU-, PDF: https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/eddea82ad2755b24c4e168c5fc2ebd40-Paper-round2.pdf.

[14] Goel, A., Pasricha, P., Kanniainen, J. (2024), *Time-series foundation model for value-at-risk forecasting*. arXiv preprint arXiv:2410.11773, available at: https://arxiv.org/abs/2410.11773.

[15] Goel, A., Pasricha, P., Magris, M., Kanniainen, J. (2025), *Foundation time-series AI model for realized volatility forecasting*. arXiv preprint arXiv:2505.11163, available at: https://arxiv.org/abs/2505.11163.

[16] Hansen, B.E. (1994), *Autoregressive conditional density estimation. International Economic Review*, 35(3), 705-730, available at: https://ideas.repec.org/a/ ier/iecrev /v35y1994i3p705-30.html, PDF: https://www.ssc.wisc.edu/~bhansen/papers/ ier_94 .pdf.

[17] Kupiec, P.H. (1995), *Techniques for verifying the accuracy of risk measurement models. Journal of Derivatives*, 3(2), 73-84, available at: https://www.pm-research.com/content/iijderiv/3/2/73.

[18] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., Wasserman, L. (2018), *Distribution-free predictive inference for regression. Journal of the American Statistical Association*, 113(523), 1094–1111, https://doi.org/10.1080/01621459.2017.1307116.

[19] Liu, X., Liu, J., Woo, G., Aksu, T., Liang, Y., Zimmermann, R., Liu, C., Savarese, S., Xiong, C., Sahoo, D. (2024), *Moirai-MoE: Empowering time series foundation models with sparse mixture of experts*. arXiv preprint arXiv:2410.10469, available at: https://arxiv.org/abs/2410.10469.

[20] Liu, C., Aksu, T., Liu, J., Liu, X., Yan, H., Pham, Q., Sahoo, D., Xiong, C., Savarese, S., Li, J. (2025), *Moirai 2.0: When less is more for time series forecasting*. arXiv preprint arXiv:2511.11698, available at: https://arxiv.org/abs/2511.11698.

[21] Marconi, B.A. (2025), T*ime series foundation models for multivariate financial time series forecasting*. arXiv preprint arXiv:2507.07296, available at: https://arxiv.org/abs/2507.07296.

[22] McNeil, A.J., Frey, R. (2000), *Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. Journal of Empirical Finance*, 7(3-4), 271-300, https://doi.org/10.1016/S0927-5398(00)00012-8.

[23] Pele, D.T., Conda, A.I., Bag, R., Mazurencu-Marinescu-Pele, M., Strat, V.A. (2023), *Financial risk meter for the Romanian stock market. Romanian Journal of Economic Forecasting*, 26(1), 5-24, available at: https://ipe.ro/new/rjef/rjef1_2023/rjef1_2023p5-24.pdf.

[24] Pele, D.T., Bolovăneanu, V., Lin, M.-B., Ren, R., Ginavar, A.T., Spilak, B., Andrei, A.-V., Toma, F.-M., Lessmann, S., Härdle, W.K. (2026), *In the beginning was the Word: LLM-VaR and LLM-ES. Expert Systems with Applications*, 295, 128676, https://doi.org/10.1016/j.eswa.2025.128676.

[25] Pele, D.T., Bolovăneanu, V., Ginavar, A.T., Lessmann, S., Härdle, W.K. (2026), *Recalibrating tail risk forecasts under temporal dependence*, available at SSRN: https://ssrn.com/abstract=6757685.

[26] Rahimikia, E., Ni, H., Wang, W. (2025), *Re(Visiting) time series foundation models in finance.* arXiv preprint arXiv:2511.18578, available at: https://arxiv.org/abs/2511.18578.

[27] Romano, Y., Patterson, E., Candès, E. (2019), *Conformalized quantile regression*. In *Advances in Neural Information Processing Systems 32* — 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver Convention Center, Vancouver, British Columbia, Canada, 8-14 December 2019, available at: https://papers.nips.cc/paper/8613-conformalized-quantile-regression. arXiv version: https://arxiv.org/abs/1905.03222.

[28] Vovk, V., Gammerman, A., Shafer, G. (2005), *Algorithmic Learning in a Random World. New York: Springer*, New York, USA, https://doi.org/10.1007/b106715.

[29] Woo, G., Liu, C., Kumar, A., Xiong, C., Savarese, S., Sahoo, D. (2024), *Unified training of universal time series forecasting transformers*. In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024),* Messe Wien Exhibition Congress Center, Vienna, Austria, 21-27 July 2024. PMLR, 235, 53140-53164, available at: https://proceedings.mlr.press/v235/woo24a.html. arXiv version: https://arxiv.org/abs/2402.02592.

[30] Zhu, Z., Chen, H., Qu, Q., Chung, V. (2025), *FinCast: A foundation model for financial time-series forecasting.* In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM 2025),* Seoul, South Korea, 10-14 November 2025, https://doi.org/10.1145/3746252.3761261, arXiv version: https://arxiv.org/abs/2508.19609.

Appendix A - Abbreviation Table

Abbreviation	Full Term	Abbreviation	Full Term
ARIMA	Autoregressive Integrated Moving Average	HS	Historical Simulation
BET	Bucharest Exchange Trading Index	LSTM	Long Short-Term Memory
BUX	Budapest Stock Exchange Index	OOS	Out-of-Sample
CC	Conditional Coverage	PX	Prague Stock Exchange Index
ADF	Augmented Dickey–Fuller	FX	Foreign Exchange
AR	Autoregressive	GARCH	Generalized Autoregressive Conditional Heteroskedasticity
ARCH	Autoregressive Conditional Heteroskedasticity	GJR-GARCH	Glosten–Jagannathan–Runkle Generalized Autoregressive Conditional Heteroskedasticity
ARIMA	Autoregressive Integrated Moving Average	HS	Historical Simulation
ARIMA-CP	ARIMA with Conformal Prediction	LLM	Large Language Model
BET	Bucharest Exchange Trading Index	LSTM	Long Short-Term Memory
BUX	Budapest Stock Exchange Index	LSTM-CP	LSTM with Conformal Prediction
CC	Conditional Coverage Test	Moirai	Salesforce Time Series Foundation Model
CEE	Central and Eastern Europe	PX	Prague Stock Index
Chronos-2	Amazon Time Series Foundation Model	SOFIX	Sofia Stock Exchange Index
Conf	Conformalized / Conformal Calibration	TimesFM	Google Time Series Foundation Model
CP	Conformal Prediction	TSFM	Time Series Foundation Model
DM	Diebold–Mariano Test	UC	Unconditional Coverage Test
ES	Expected Shortfall	VaR	Value-at-Risk
FM	Foundation Model	WIG20	Warsaw Stock Exchange Index
FRTB	Fundamental Review of the Trading Book	Z_2	Acerbi–Szekely Expected Shortfall Backtest Statistic

Appendix B - Diagnostic Analysis

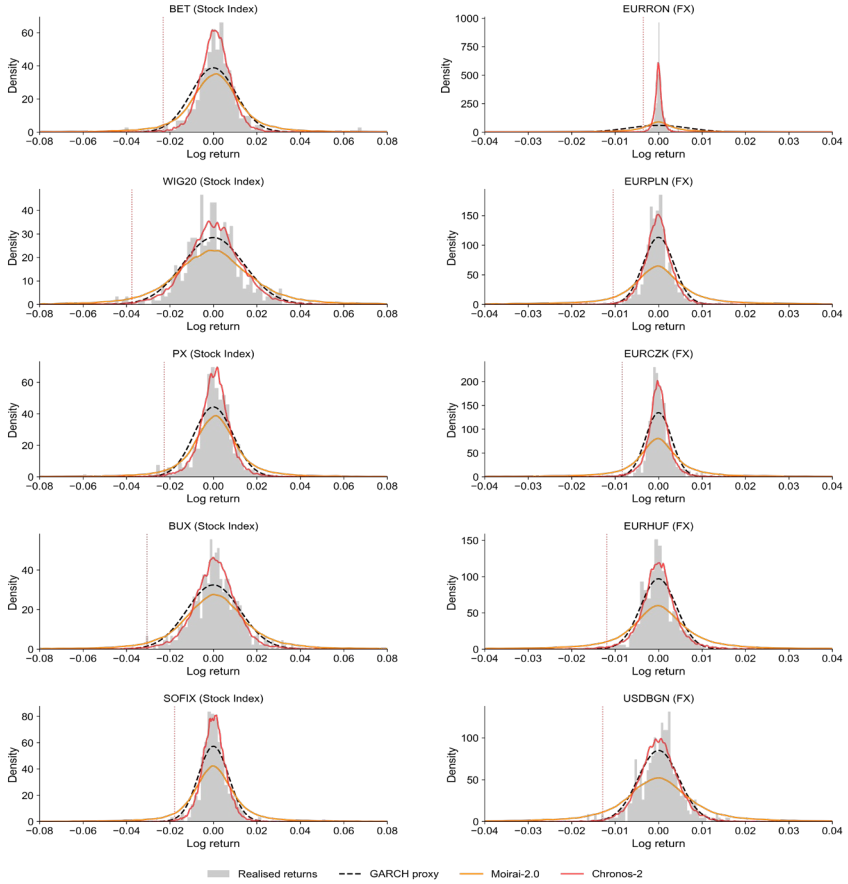

Figure A1. Predictive Distribution Comparison across all 10 CEE Series: Realized Returns vs GARCH proxy, Moirai-2.0, TimesFM-2.5 and Chronos-2
Source: Authors' own creation.

Figure A1 presents a comparison between the predictive distributions generated by the three foundation models and the realized return distributions across all ten CEE series, with stock indices shown in the left column and foreign exchange pairs in the right column. All three foundation models produce predictive distributions that broadly capture the shape and dispersion of realized returns. The contrast between equity and foreign exchange returns is informative: foreign exchange returns exhibit a narrower support relative to equities, and all three TSFMs appropriately adjust their predictive dispersion across asset classes. From a quantitative perspective, the average conformal adjustment increases the magnitude of VaR estimates by approximately 38% for TimesFM 2.5 and 53% for Moirai 2.0, while Chronos-2 requires a more modest adjustment, consistent with its already competitive raw performance.

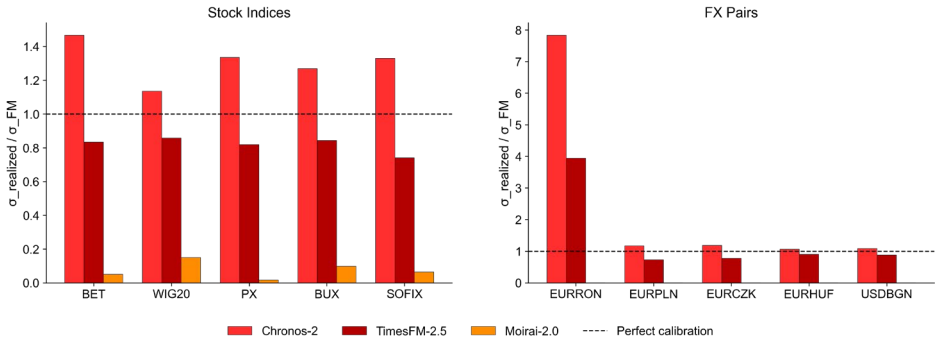

Figure A2. Dispersion Ratios (realized vs model-implied volatility) across Stock Indices and FX Pairs

Source: Authors' own creation.

Furthermore, Figure A2 illustrates the dispersion ratios between realized and implied volatilities. All three foundation models capture the general volatility dynamics, though some heterogeneity remains across markets. Figures A3 and A4 further illustrate the improvement introduced by the conformal wrapper.

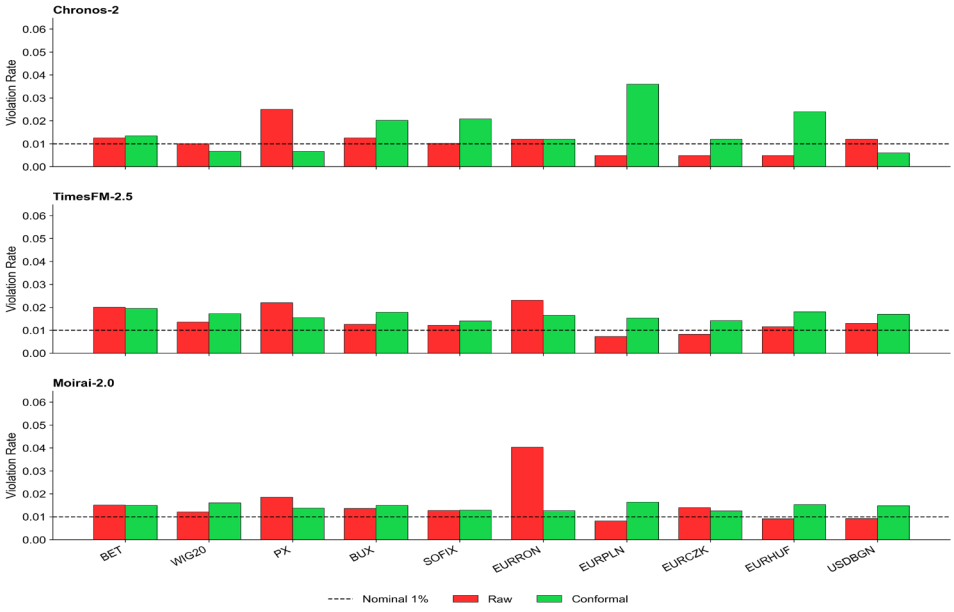

Figure A3. Conformal Recalibration: Raw vs Conformal Violation Rates across all 10 CEE Series, per Foundation Model

Source: Authors' own creation.

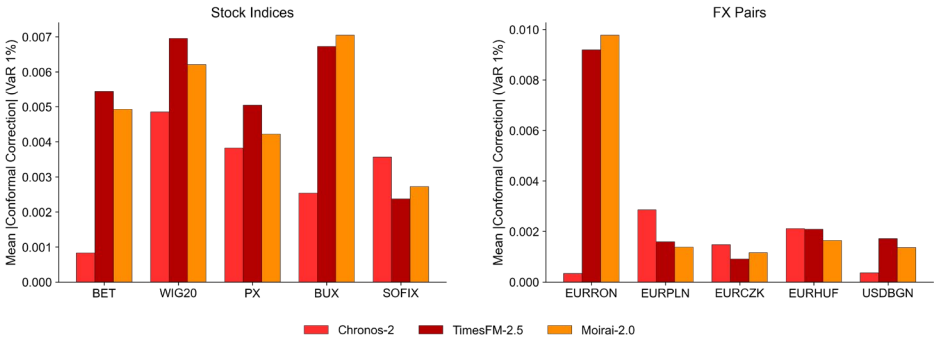

Figure A4. Conformal Correction Magnitude across Stock Indices, FX Pairs, and Foundation Models

Source: Authors' own creation.