*modelling*

MDPI

*Article*

# A Reproducible and Regime-Aware SARIMA Modelling Framework for National Air Traffic Forecasting: Evidence from Türkiye (2018–2025)

**Recep Kaş [1,*], Mehmet Şen [2], Seda Arık Hatipoğlu [3] and Mehmet Konar [3]**

[1] Department of Aircraft Technology, Cappadocia University, 50420 Nevşehir, Türkiye
[2] Distance Education Application and Research Center, Social Sciences University of Ankara, 06050 Ankara, Türkiye; mehmet.sen@asbu.edu.tr
[3] Department of Aviation Electrical and Electronics, Faculty of Aeronautics and Astronautics, Erciyes University, 38260 Kayseri, Türkiye; arikseda@erciyes.edu.tr (S.A.H.); mkonar@erciyes.edu.tr (M.K.)
* Correspondence: recep.kas@kapadokya.edu.tr

**Abstract**

Reliable short-term air traffic forecasts are important for operational planning in national airspace systems. This study develops a transparent forecasting framework for Türkiye's monthly aircraft movements using publicly available data from the General Directorate of State Airports Authority (DHMİ) for 2018–2025. Because DHMİ releases may follow cumulative within-year reporting, month-specific increments are reconstructed through within-year differencing and checked through simple audit procedures. The empirical analysis compares seasonal naïve, ETS, and a constrained SARIMA family under leakage-free evaluation, combining a strict 2025 holdout with expanding-window rolling-origin validation. Forecast performance is assessed using standard accuracy metrics and complemented by Diebold–Mariano comparisons, which are interpreted cautiously, given the short holdout length. To examine instability around the pandemic period, this study also reports structural-break and stability diagnostics as supportive evidence rather than definitive identification. Uncertainty is evaluated through backtested 80% and 95% prediction intervals, comparing nominal SARIMA intervals, parametric bootstrap, split conformal prediction, and adaptive conformal inference (ACI). The results show that SARIMA provides the strongest point-forecast performance among the benchmarked models, while adaptive conformal calibration offers a useful balance between empirical coverage and interval width under changing conditions. Overall, this study provides a reproducible and operationally interpretable baseline for national air traffic forecasting in Türkiye and a clear benchmark for future multivariate extensions.

**Keywords:** air traffic forecasting; SARIMA; rolling-origin; cross-validation; Diebold–Mariano test; structural break; prediction intervals; bootstrap; conformal prediction; adaptive conformal inference



Academic Editor: Wei Gao

Received: 18 February 2026
Revised: 8 April 2026
Accepted: 8 April 2026
Published: 21 April 2026



## 1. Introduction

Air traffic forecasting is a core capability for ATM and airport–airspace operations, informing staffing, sector configuration, slot coordination, and demand–capacity balancing. At the national scale, traffic series typically exhibit strong annual seasonality and sensitivity to shocks that can alter both demand levels and volatility. In Türkiye, national traffic reflects not only airport movements but also substantial transit/overflight flows, making

the system particularly exposed to regime changes (e.g., pandemic disruptions) and to shifts in variability as well as level. Operational aviation forecasting studies often begin with transparent univariate baselines that are easy to audit and provide strong performance for seasonal monthly series [1].

Air traffic demand is shaped by macroeconomic conditions, tourism dynamics, network connectivity, and policy/regulatory constraints; therefore, multivariate and exogenous-driver models (e.g., SARIMAX and econometric-driver specifications) are commonly explored once a stable univariate baseline has been established. SARIMAX-style models are a natural next step when covariates such as tourism indicators, weather disruptions, or event calendars are available and reliably measured. However, practical gains can be heterogeneous across locations and horizons, and adding drivers can introduce governance complexity (data availability, revisions, and drift), reinforcing the value of a reproducible univariate baseline as a "reference point" before more complex system designs are adopted.

Machine learning, deep learning, and hybrid approaches have become increasingly prominent in recent aviation forecasting research, with recurrent neural networks—most notably Long Short-Term Memory (LSTM)—often used to model nonlinear temporal dependencies in passenger-demand and passenger-flow series [2]. Comparative evidence in airport passenger-flow settings suggests that neural models (including RNN/LSTM variants) can outperform classical seasonal baselines under certain conditions, but reported gains are highly contingent on feature construction, training protocol, and evaluation design, and may deteriorate under volatile regimes or distribution shifts if validation is not carefully aligned with operational realities [3]. In closely related demand-forecasting domains such as tourism, hybrid architectures that couple a SARIMA-like seasonal structure with deep components (e.g., SARIMA–CNN–LSTM) have been proposed to combine interpretability from linear seasonal modeling with improved nonlinear pattern capture [4]. These developments motivate two operational principles for ATM deployment: (i) maintain a transparent baseline model to preserve auditability and provide a stable reference for decision-making, and (ii) evaluate advanced models under strict time-series protocols (e.g., rolling-origin/blocked evaluation) to avoid overly optimistic conclusions that can arise when temporal dependence is not respected [5].

Leakage-free evaluation and statistically defensible comparisons are essential to close a recurring methodological gap in applied forecasting: the mismatch between reported accuracy and deployable reliability. Because operational forecasts are generated strictly forward in time, time-series settings require leakage-free evaluation designs; random train–test splits and conventional k-fold cross-validation can produce overly optimistic performance estimates, whereas expanding-window and rolling-origin validation more faithfully reflect how forecasts are issued in practice [1]. Beyond aggregate error metrics, operational stakeholders frequently require statistical evidence that observed performance differences are substantive rather than incidental. In this context, forecast-comparison procedures such as the Diebold–Mariano test provide a widely used and defensible framework for assessing whether accuracy improvements over baselines are statistically significant under alternative loss functions [6].

Regime change, structural breaks, and diagnostic checking represent a second major challenge in operational aviation forecasting, where non-stationarity can arise abruptly and undermine stable-parameter assumptions. Major disruptions may introduce structural breaks that invalidate classical inference and degrade model performance; accordingly, break and stability tests—such as Chow-type procedures for a prespecified breakpoint and CUSUM-style tests for parameter constancy—provide formal evidence of instability and motivate regime-aware evaluation designs and re-training policies [7]. Complemen-

tary diagnostic checking remains critical for assessing model adequacy and calibrating uncertainty, including residual autocorrelation tests [8], normality checks for parametric assumptions [9], and conditional heteroskedasticity tests capturing volatility clustering [10]. These tools are especially relevant for aviation series in which shocks can alter both the level and the variance dynamics, making routine stability assessment and post-fit diagnostics essential parts of an operational modeling workflow.

Uncertainty quantification and conformal prediction under dependence and shift are increasingly treated as first-class requirements for decision support, yet many applied forecasting studies still report nominal prediction intervals without empirically validating their coverage. Classical conformal prediction provides finite-sample, distribution-free marginal coverage guarantees under an exchangeability assumption, but temporal dependence and a distribution shift in time series typically violate exchangeability and can erode nominal validity if used naively [11,12]. Recent work has therefore proposed conformal variants designed specifically for dynamic settings. In particular, EnbPI constructs sequential prediction intervals by wrapping an ensemble predictor and updating conformity scores online, and it is empirically robust in dependent time-series regimes without relying on standard exchangeability conditions [12]. More directly addressing the distribution shift, adaptive conformal inference (ACI) performs online calibration to target a desired coverage frequency over long horizons even when the data-generating process varies arbitrarily over time, making it attractive for operational environments characterized by regime changes [13]. Together, these approaches motivate reporting uncertainty as a backtested property—coverage and sharpness—rather than as a purely nominal statement.

Recent post-pandemic aviation forecasting studies in Europe emphasize regime-aware modeling and distribution-shift robustness, particularly in ATM contexts where corridor reconfiguration and traffic redistribution affect national FIR volumes. Incorporating stability diagnostics and backtested uncertainty aligns this study with operational forecasting practices reported in European ATM modernization initiatives.

This study provides a transparent baseline for forecasting Türkiye's monthly national aircraft movements using public DHMİ statistics. Specifically, it (i) reconstructs month-specific increments from cumulative reporting, (ii) compares seasonal naïve, ETS, and constrained SARIMA models under leakage-free evaluation, (iii) reports supportive evidence of regime instability around the pandemic period, and (iv) backtests alternative prediction-interval methods on a strict 2025 holdout. The aim is not to claim a fully explanatory forecasting system, but to establish a reproducible and operationally interpretable benchmark that can support future multivariate and hybrid extensions.

## 2. Materials and Methods

This study follows a sequential forecasting workflow that is summarized in Figure 1. The pipeline consists of eight steps: (1) acquisition of monthly DHMİ tables for 2018–2025; (2) reconstruction of month-specific increments from cumulative within-year reporting; (3) simple audit checks for missing values, negative increments, year-boundary consistency, and annual reconciliation; (4) estimation of benchmark models, including seasonal naïve, ETS, and a constrained SARIMA family; (5) leakage-free evaluation using a strict 2025 holdout and expanding-window rolling-origin validation; (6) forecast comparison using standard accuracy metrics and Diebold–Mariano tests; (7) regime-instability assessment via Chow and CUSUM diagnostics; and (8) backtesting of 80% and 95% prediction intervals using nominal, bootstrap, split conformal, and adaptive conformal methods. The purpose of this design is to provide a transparent and operationally interpretable benchmark rather than a fully explanatory multivariate forecasting system.

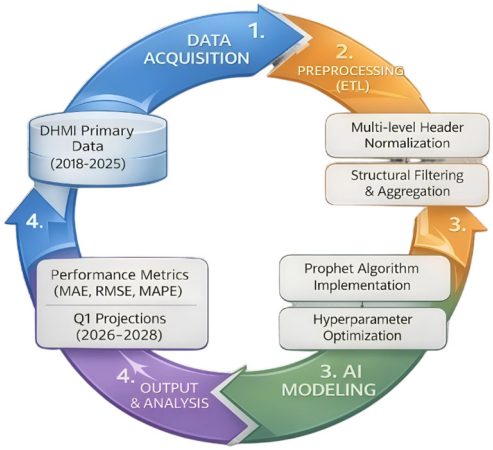


**Figure 1.** Research framework and end-to-end pipeline.

### *2.1. Data Source, Study Variables, and Monthly Series Construction*

We used Türkiye's national monthly aircraft movement statistics, published by the General Directorate of State Airports Authority (DHMİ). The raw monthly tables ("Türkiye Geneli Havalimanı İstatistikleri" → monthly "Toplam Uçak" spreadsheets) were downloaded from the public DHMİ portal on 2 February 2026 for the period January 2018–December 2025 (96 calendar months). The primary forecasting target was the national total number of aircraft movements per month (including overflights), as it directly supports demand–capacity planning and reflects both airport activity and en-route traffic at the national scale. When an overflight series was provided separately, we computed the overflight share as a descriptive indicator:

$$s_t = \frac{\text{Overflight}_t}{\text{Total}_t} \tag{1}$$

where $\text{Overflight}_t$ denotes overflight movements and $\text{Total}_t$ denotes total movements in month $t$. We used $s_t$ for interpretation and diagnostics (e.g., separating corridor-driven changes from airport-driven changes), but we did not use it as an exogenous regressor in the baseline models so as to keep the reference pipeline strictly univariate and audit-friendly.

The DHMİ notes that monthly releases may be reported as cumulative totals within a calendar year. To ensure a consistent month-by-month series for forecasting, we reconstructed monthly increments via within-year differencing. Let $C_{y,m}$ denote the reported value for year $y$ and month $m$. We defined the reconstructed monthly increment $x_{y,m}$ as:

$$y_t = C_t - C_{t-1} \text{ for } t = \text{Feb}, \ldots, \text{Dec}, \tag{2}$$

$$y_{\text{Jan}} = C_{\text{Jan}} \tag{3}$$

The resulting series $\{x_t\}$ is the sole input used for all forecasting, leakage-free evaluations, and uncertainty backtesting.

To reduce the risk of silent pipeline failures, we applied a small set of deliberately simple, loggable checks prior to modeling: (i) non-negativity of reconstructed increments; (ii) year-boundary reset enforcement (differencing strictly within the year); (iii) missing-month detection with explicit handling rules (exclude vs. impute, always logged); (iv) outlier/spike flagging for manual review; and (v) annual total reconciliation when annual totals are available (the sum of reconstructed months versus the reported annual total).

### 2.2. Benchmark Models

We benchmarked seasonal SARIMA against two practical baselines that are widely used and strong comparators for monthly seasonal time series, ensuring that any improvement from more complex models was meaningful rather than marginal.

We first used the seasonal naïve baseline with seasonal period $s = 12$, reflecting the annual seasonality in monthly data. The $h$-step-ahead forecast at time $t$ is defined as

$$\hat{y}_{t+h|t} = y_{t+h-s}, s = 12 \quad (4)$$

i.e., the forecast for a future month is set equal to the observed value from the same calendar month one year earlier. Here, $y_t$ denotes the observed series and $\hat{y}_{t+h|t}$ denotes the forecast issued at time $t$ for horizon $h$. This baseline is intentionally simple yet often competitive when annual seasonality dominates, and it provides a transparent "no-model" reference that any advanced method should outperform.

As a second baseline, we adopted ETS (Holt–Winters) exponential smoothing with additive seasonality, a standard and interpretable approach that admits a state-space formulation and supports likelihood-based estimation and prediction intervals [14]. We used an additive seasonal form because the target is measured in counts and the dominant seasonal pattern is annual; additive seasonality is appropriate when seasonal fluctuations are reasonably stable in absolute magnitude rather than scaling proportionally with the series level. In practice, ETS represents the series through components for the level (and optionally trend) and a repeating seasonal term and extrapolates these components to produce forecasts.

These baselines were selected because they are operationally familiar, easy to audit, and sufficiently strong, and improvements over them can be interpreted as practically meaningful.

### 2.3. SARIMA Candidate Family

We used a seasonal ARIMA (SARIMA) methodology as a transparent reference standard for monthly seasonal time series. SARIMA models remain a cornerstone in applied forecasting because they are interpretable and explicitly represent both seasonal differencing and seasonal autocorrelation structures [15]. In this study, SARIMA is positioned as a governance-grade baseline: it is flexible enough to capture short-term and seasonal dynamics, yet it is structured enough to be auditable and reproducible.

To define the SARIMA candidate family, we assumed a monthly seasonal period of $s = 12$, where $s$ denotes the number of observations in one seasonal cycle; in monthly data, $s = 12$ corresponds to one year [15]. We fit a constrained SARIMA family to balance robustness and parsimony in a small-sample monthly setting. Specifically, each candidate model is indexed by two parameter triplets, $(p, d, q)$ and $(P, D, Q)$, written compactly as

$$(p, d, q) \times (P, D, Q)_{S=12} \quad (5)$$

Here, $p$ is the non-seasonal autoregressive (AR) order, $q$ is the non-seasonal moving-average (MA) order, and $d$ is the non-seasonal differencing order; likewise, $P$ and $Q$ denote the seasonal AR and MA orders at lag $s$, and $D$ denotes the seasonal differencing order [15]. Candidate specifications are restricted to the grid,

$$p, q \in \{0, 1, 2\}, P, Q \in \{0, 1\}, d \in \{0, 1\}, D \in \{0, 1\}, s = 12 \quad (6)$$

and a constant (intercept) term is included when compatible with the chosen differencing orders. Importantly, all model fitting and selection are performed within each training win-

dow only, preserving a leakage-free evaluation that is consistent with standard forecasting practice [1].

We intentionally restricted the candidate grid because the objective was a reproducible reference baseline rather than an unrestricted search. Constraining the family reduces overfitting risk and improves stability across rolling training windows while still allowing sufficient flexibility to represent short-term dynamics and annual seasonality [15].

Models are estimated via maximum likelihood, and candidate selection within each training window is based on AIC. Because information criteria can occasionally favor numerically fragile models in small samples, we applied three explicit plausibility gates before accepting the selected candidate. (i) Convergence gate: the optimizer must report successful convergence, and the estimated parameter covariance must be finite; otherwise, the candidate is rejected. (ii) Stability gate: we enforced stationarity/invertibility by requiring all AR and MA roots to lie outside the unit circle with a small margin

$$|\, r \,| \;>\; 1 + \varepsilon \tag{7}$$

We set $\varepsilon$; near-unit-root fits are discarded to avoid unstable dynamics. (iii) Forecast sanity gate: we generated a fixed-horizon multi-step forecast H from the candidate and rejected models whose forecasts exhibit implausible explosive behavior relative to the historical scale. We logged the number of discarded candidates per training window and reported it to make model selection auditable.

### *2.4. Evaluation Design: Strict Holdout and Expanding-Window Rolling-Origin Cross-Validation*

Forecast evaluation in time series differs fundamentally from i.i.d. machine learning because observations are time-ordered and typically serially dependent. As a result, random train–test splits and standard $k$-fold cross-validation can leak future information into the training process and yield overly optimistic performance estimates. For deployment realism, evaluation should therefore respect temporal ordering and use rolling-origin designs that mimic how forecasts are produced in operations [16].

We employed two complementary leakage-free evaluation designs. First, we used a strict holdout (deployment-style test) in which models are trained on January 2018–December 2024 and evaluated on the full calendar year 2025 (12 months). This provides a clean "future" test year for final performance reporting and for uncertainty backtesting. Within the 2025 test year, all accuracy and interval backtests were computed from rolling one-step-ahead forecasts: for each test month $t$, the model was refitted using all data available up to $t-1$, and then a forecast was issued for month $t$. This procedure avoids look-ahead and matches operational monthly updating.

Second, we used expanding-window rolling-origin cross-validation (RO-CV) to assess generalization repeatedly as the training history grew. Specifically, we evaluated annual test blocks by training through 2021 and testing in 2022, training through 2022 and testing in 2023, training through 2023 and testing in 2024, and training through 2024 and testing in 2025. At each origin, the model is refitted on all available historical data and then used to forecast the next block, closely mirroring real operational practice and the standard rolling-origin analog of cross-validation in time-series forecasting [16].

For planning-oriented projections beyond the evaluation window, we also report fixed-origin multi-step forecasts: forecasts for 2026 Q1 (January–March 2026) are generated from the final model fitted through December 2025, producing horizons $h = 1, 2, 3$ together with corresponding 80% and 95% prediction intervals. Across all evaluation segments, we report MAE, RMSE, MAPE, and sMAPE, where MAE denotes the mean absolute error, RMSE denotes the root mean squared error, MAPE denotes the mean absolute percentage error, and sMAPE denotes the symmetric mean absolute percentage error [17].

### 2.5. Accuracy Metrics

Forecast accuracy is evaluated using four standard measures: MAE, RMSE, MAPE, and sMAPE. These metrics provide complementary views of forecast performance in both absolute and relative terms and are widely used in time-series forecasting studies. Let $\hat{y}_t$ denote the observed value $\hat{y}_t$ the forecast, and $e_t = y_t - \hat{y}_t$ the forecast error for $t = 1, \ldots, n$. The metrics are computed as:

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |e_t| \tag{8}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \tag{9}$$

$$MAPE = \frac{100}{n}\sum_{t=1}^{n} \left| \frac{e_t}{y_t} \right| \tag{10}$$

$$sMAPE = \frac{100}{n}\sum_{t=1}^{n} \frac{2|e_t|}{|y_t| + |\hat{y}_t|} \tag{11}$$

To make the comparison explicit, let $e_t^{(A)} = y_t - \hat{y}_t^{(A)}$ and $e_t^{(B)} = y_t - \hat{y}_t^{(B)}$ denote the forecast errors of two competing models $A$ and $B$ at month $t$. We defined a loss function $L(\cdot)$ that maps an error to a nonnegative penalty, and we considered two standard choices aligned with our metric reporting: squared-error loss $L(e) = e^2$ and absolute-error loss $L(e) = |e|$. For each month, we computed the loss differential

$$d_t = L\left(e_t^{(A)}\right) - L\left(e_t^{(B)}\right) \tag{12}$$

where $d_t < 0$ indicates that model $A$ has lower loss (better accuracy) than model $B$ at time $t$ under the selected loss.

The Diebold–Mariano (DM) test evaluates whether the average loss differential is zero:

$$H_0: \ \mathbb{E}[d_t] = 0 \text{ vs. } H_1: \ \mathbb{E}[d_t] \neq 0 \tag{13}$$

where $\mathbb{E}[\cdot]$ denotes expectation over time. In practice, the sample mean differential is

$$\bar{d} = \frac{1}{n}\sum_{t=1}^{n} d_t \tag{14}$$

with $n$ denoting the number of forecast–observation pairs in the evaluation segment (e.g., $n = 12$ for the 2025 holdout). The DM statistic standardizes $\bar{d}$ by an estimate of its standard error:

$$DM = \frac{\bar{d}}{\sqrt{\widehat{Var}\left(\bar{d}\right)}} \tag{15}$$

$\widehat{\mathrm{Var}}\left(\bar{d}\right)\{d_t\}p$ where $\mathrm{V_{hat}}$ is computed using a heteroskedasticity-and-autocorrelation-consistent (HAC) estimate based on the autocovariance of d_t, acknowledging that loss differentials may be serially correlated in time-series settings [6]. Because the strict holdout contains only 12 monthly observations, we apply the small-sample correction of [18,19] and treat the DM test as a supplementary inferential check rather than as the primary basis for model ranking.

$\bar{d}$, in reporting a negative DM statistic under the chosen loss, is interpreted as being consistent with a lower loss for SARIMA relative to the benchmark. However, statistical

significance on the short 2025 holdout should be interpreted cautiously; the rolling-origin evidence is therefore emphasized as the primary basis for comparative assessment.

### *2.6. Structural Break and Regime Instability Evidence*

Aviation time series are often exposed to abrupt disruptions, recovery phases, and evolving route structures that can violate constant-parameter assumptions. To assess whether Türkiye's monthly national aircraft movements exhibit instability around the pandemic period, we report two standard diagnostics of structural change and parameter constancy. These tests are used as supportive evidence of instability, not as a standalone identification strategy.

First, we applied a Chow test at the prespecified breakpoint of April 2020, motivated by the timing of the COVID-19 shock [7]. Because this breakpoint is chosen on substantive grounds rather than discovered through a data-driven break search, the Chow result is interpreted as a targeted diagnostic rather than as general proof of a uniquely identified break date. Second, we implemented the CUSUM stability test based on recursive residuals to assess broader parameter constancy over time without requiring a single fixed breakpoint [14]. In operational terms, CUSUM is useful as a monitoring-oriented diagnostic because it can detect gradual drift as well as abrupt changes.

Taken together, these procedures indicate whether the series can reasonably be treated as generated by one stable regime over 2018–2025. Evidence of instability motivates two design choices adopted in this study: emphasizing rolling-origin evaluation across changing conditions and backtesting interval coverage empirically rather than relying only on nominal parametric assumptions [20].

### *2.7. Residual Diagnostics*

Residual diagnostics are used to assess whether a fitted model has captured the main temporal structure in the data and to contextualize the credibility of parametric uncertainty statements. In this study, we apply three standard families of residual checks to the fitted SARIMA models, focusing on properties that directly affect forecast adequacy and the interpretation of prediction intervals.

First, to evaluate autocorrelation and lack-of-fit, we apply the Ljung–Box $Q$-test at lags 12 and 24, corresponding to one- and two-year horizons in monthly data [8]. The purpose is to test whether statistically detectable serial dependence remains in the residuals after model fitting. Persistent residual autocorrelation indicates that the model may be underspecified (e.g., missing seasonal or short-term dynamics), and the Ljung–Box framework is a standard portmanteau approach in ARIMA settings [8,15].

Second, we examine normality using the Jarque–Bera test [9]. This check is relevant because nominal SARIMA prediction intervals are typically derived under Gaussian error assumptions; substantial departures from normality can lead to miscalibrated parametric intervals even when point forecasts are reasonable. Normality testing, therefore, helps determine when it is prudent to rely more heavily on empirical or nonparametric interval methods [21,22].

Third, we assess conditional heteroskedasticity by testing for ARCH effects using Engle's ARCH–LM procedure [10]. Volatility clustering and time-varying variance are common in disrupted or rapidly changing systems, including aviation demand, where shocks can alter variance dynamics even if the mean structure is well captured. Evidence of ARCH behavior motivates caution in interpreting homoskedastic Gaussian intervals and supports the use of bootstrap- or conformal-style uncertainty approaches that are more robust to variance shifts [23].

Overall, these diagnostics are not intended to "validate" a model in an absolute sense. Rather, they clarify which modeling assumptions are most tenuous in this dataset and, in turn, which interval construction methods are more likely to be conservative, misleading, or preferable in deployment settings.

### *2.8. Prediction Intervals: Why Backtesting Is Required*

In operational ATM planning, decision-makers need uncertainty information in addition to point forecasts, because capacity planning and risk management depend on how wrong a forecast could be, not only on its central value. Because prediction intervals can be miscalibrated under regime changes, we evaluated interval quality using leakage-free one-step-ahead backtesting on the 2025 holdout. For this reason, we treated interval quality as an empirical property and evaluated prediction intervals via leakage-free, one-step-ahead backtesting on the strict 2025 holdout [24].

For each month $t$, the model produces a prediction interval $[L_t, U_t]$ for the next observation $y_t$. Here, $L_t$ is the lower bound and $U_t$ is the upper bound, both issued using only information available up to $t-1$. We assess interval performance by using three complementary quantities: coverage, sharpness, and a single combined proper scoring rule.

Coverage measures reliability: this is defined as the fraction of test months for which the realized value $y_t$ lies inside the predicted interval:

$$\text{Coverage} = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{L_t \le y_t \le U_t\} \tag{16}$$

where $1\{\cdot\}$ is an indicator function that equals 1 when the condition is true and 0 otherwise, and $n$ is the number of evaluated months (e.g., $n = 12$ for the 2025 holdout). For a nominal 95% interval, we expect empirical coverage close to 0.95; large deviations indicate under- or over-coverage.

Sharpness measures informativeness: A perfectly reliable interval is not useful if it is extremely wide. We therefore report the interval width,

$$\text{Width}_t = U_t - L_t, \tag{17}$$

and summarize it across the test period (e.g., by the mean or median width). Smaller widths indicate sharper, more informative intervals, but only if coverage remains close to the nominal level.

To summarize calibration and sharpness in a single interpretable quantity, we also compute the interval score (Winkler score), a proper scoring rule commonly used in probabilistic forecasting. For a nominal level $1-\alpha$ (e.g., $\alpha = 0.2$ for 80% intervals and $\alpha = 0.05$ for 95% intervals), the interval score for observation $y_t$ and interval $[L_t, U_t]$ is:

$$IS_\alpha(L_t, U_t; y_t) = (U_t - L_t) + \frac{2}{\alpha}(L_t - y_t)\,\mathbf{1}\{y_t < L_t\} + \frac{2}{\alpha}(y_t - U_t)\,\mathbf{1}\{y_t > U_t\} \tag{18}$$

Each term has a clear meaning:

- $(U_t - L_t)$ is the baseline width term: narrower intervals are rewarded because they provide more precise uncertainty information.
- $1\{y_t < L_t\}$ is 1 only when the observation falls below the lower bound, i.e., the interval misses on the low side.
- $1\{y_t > U_t\}$ is 1 only when the observation falls above the upper bound, i.e., the interval misses on the high side.
- $\frac{2}{\alpha}(L_t - y_t)1\{y_t < L_t\}$ is the penalty for under-coverage on the low side: the further $y_t$ is below $L_t$, the larger the penalty. The factor $2/\alpha$ makes this penalty stronger when

$\alpha$ is smaller (e.g., 95% intervals should miss less often, so misses are penalized more heavily).

- $\frac{2}{\alpha}(y_t - U_t)1\{y_t > U_t\}$ is the penalty for under-coverage on the high side, defined analogously.

Lower values of $IS_\alpha$ indicate better interval performance because they correspond to intervals that are both narrow and rarely miss the true value; the score increases when intervals are wide or when they fail to cover the observation.

All interval results are computed under a leakage-free, deployment-style procedure. Within the 2025 holdout, we generate rolling one-step-ahead intervals: at each month $t$, models are refit using data up to $t-1$, an interval $[L_t, U_t]$ is produced for month $t$, and coverage, width, and interval score are updated sequentially. This design ensures that interval backtests reflect how uncertainty would be produced and consumed in real operations [25].

### *2.9. Prediction Interval Methods Compared (Nominal, Bootstrap, Conformal, and Adaptive Conformal)*

To characterize uncertainty under increasing robustness to misspecification and distribution shift, we compare four prediction-interval (PI) constructions. All methods are evaluated under the same leakage-free, rolling one-step-ahead protocol on the 2025 holdout; at each test month $t$, the forecasting model is refit using data available up to $t-1$, a point forecast $\hat{y}_t$ is produced, and an interval $[L_t, U_t]$ is generated using one of the methods below.

Nominal SARIMA prediction intervals: The first approach uses the standard model-based intervals implied by the fitted SARIMA model. These intervals are derived from the estimated innovation variance and the model's multi-step forecast error variance under the usual assumptions of approximate Gaussian innovations and correct model specification. They are computationally efficient and serve as a natural baseline for probabilistic forecasting; however, they can be miscalibrated when assumptions are violated—particularly under structural breaks, heavy tails, or non-constant variance [26].

Parametric bootstrap prediction intervals: The second approach uses a parametric bootstrap to approximate forecast uncertainty by simulation. In each training window, we simulate multiple future paths from the fitted SARIMA model by repeatedly drawing innovations from the estimated error distribution (typically Gaussian with fitted variance) and propagating the SARIMA dynamics forward. The empirical quantiles of the simulated forecast distribution define the interval bounds. Compared with purely nominal formulas, bootstrap intervals can better reflect parameter-estimation uncertainty and non-ideal residual behavior while remaining relatively straightforward to implement for ARIMA-type models [27].

Split conformal prediction intervals: The third approach applies conformal inference to obtain distribution-free marginal coverage under exchangeability [11]. For computational simplicity, we adopt a split conformal design that separates model fitting from calibration: we fit the forecasting model on a training period, compute nonconformity scores on a held-out calibration period, and then construct prediction intervals by inflating the point forecast with an appropriate quantile of the calibration scores. Concretely, for one-step-ahead forecasts, we define the nonconformity score as the absolute residual

$$s_t = | y_t - \hat{y}_t | \tag{19}$$

Let $q_{1-\alpha}$ denote the $(1-\alpha)$-quantile of $\{s_t\}$ computed over the calibration set. The split conformal PI for level $1-\alpha$ is then:

$$[L_t, U_t] = [\hat{y}_t - q_{1-\alpha},\ \hat{y}_t + q_{1-\alpha}] \tag{20}$$

In our implementation, the calibration window was fixed to 2023–2024, and an evaluation was conducted in 2025. This design yields only 24 monthly calibration observations, so that the resulting conformal quantile estimates may be unstable in a small sample [28]. We therefore treat split conformal primarily as a simple reference method rather than as a definitive uncertainty solution in this application. Because strict exchangeability is typically violated in time series, split conformal can under-cover under temporal dependence or distribution shift, which further motivates adaptive variants.

$W = 24$ $\gamma = 0.15$ Adaptive/rolling conformal inference (ACI): The fourth approach targets the distribution shift by updating calibration online so that empirical coverage tracks the desired level over time [13]. We implemented ACI by using a rolling calibration window of size W months and an adaptation rate gamma. Intuitively, the method monitors recent miscoverage and adjusts the effective quantile used to set the conformal half-width, widening intervals when recent coverage is too low and narrowing them when coverage is overly conservative. Even so, ACI should be interpreted pragmatically: with a short evaluation horizon, its apparent calibration gains remain application-specific rather than conclusive [29].

Finally, we position our uncertainty evaluation relative to time-series-specific conformal approaches such as EnbPI, which constructs sequential intervals for dependent time series without relying on strict exchangeability [12]. This broader literature motivates our emphasis on backtested coverage and sharpness rather than on purely nominal interval claims [30].

### *2.10. Operational Assurance Checklist*

Because the objective of this study is not only methodological benchmarking but also deployment readiness, we summarized a minimal operational assurance checklist that links the modeling choices above to practical governance controls. First, data and code versioning are required to ensure that cumulative-to-monthly reconstruction, ETL steps, and downstream results can be reproduced exactly in future runs and audits. We also recommend enforcing leakage-free evaluation and continuous coverage monitoring as part of operational governance. Second, we recommend continuous coverage monitoring for prediction intervals by tracking rolling empirical coverage so that miscalibration is detected early [18]. Third, we include drift and break indicators—combining formal regime tests (e.g., Chow/CUSUM-style evidence) with simple operational proxies (e.g., sustained forecast bias or abrupt changes in residual variance)—to provide early warnings of regime changes [7,14]. Finally, we specify threshold-based alerts and human-in-the-loop escalation, so that abnormal deviations trigger review rather than silently propagating into planning decisions [31]. This assurance framing aligns with best-practice warnings in forecasting evaluation and operational analytics: improper evaluation and unmonitored degradation can produce systematically misleading conclusions even when headline accuracy appears acceptable [16].

### *2.11. Implementation and Reproducibility Details*

To ensure full reproducibility of the proposed framework, all analyses were conducted in a controlled and fully scripted computational environment. The workflow was implemented in Python (version 3.11) by using widely adopted scientific computing libraries. Time-series modeling was performed using the statsmodels library (version 0.14), including the SARIMAX class for SARIMA estimation and the ETSModel implementation for exponential smoothing. Data preprocessing, reconstruction of monthly increments, and rolling-origin evaluation were

handled using pandas (version 2.1) and NumPy (version 1.26). Figures were produced using matplotlib (version 3.8).

Seasonal naïve forecasts were generated directly using lag-12 observations. For the SARIMA benchmark, candidate models were estimated via maximum likelihood within the constrained grid defined in Section 2.3, using the statsmodels SARIMAX implementation with stationarity and invertibility constraints enforced. Model selection within each training window was based on AIC, and all refitting procedures followed the leakage-free rolling-origin design described in Section 2.4.

Parametric bootstrap prediction intervals were constructed using simulation from the fitted SARIMA model. For each forecast origin, B = 1000 bootstrap paths were generated by sampling innovations from the estimated residual distribution and recursively propagating the model forward. Prediction interval bounds were obtained from the empirical quantiles of the simulated forecast distribution.

Split conformal prediction was implemented using absolute residual nonconformity scores computed over the fixed calibration window spanning 2023–2024 (24 monthly observations). Adaptive conformal inference (ACI) employed a rolling calibration window of size W = 24 with adaptation rate $\gamma = 0.15$, which is consistent with the configuration used in the empirical evaluation. All conformal intervals were generated sequentially using the same rolling one-step-ahead protocol as the baseline models.

All evaluation procedures, including expanding-window refitting, accuracy metric computation, Diebold–Mariano testing, and prediction-interval backtesting, were implemented through deterministic scripts. Random seeds were fixed for bootstrap and conformal procedures to ensure replicability of the reported results. Together, these implementation details provide a fully reproducible specification of the proposed forecasting framework.

## 3. Results

This section reports the empirical findings from the reconstructed national monthly aircraft movement series and the forecasting evaluations described in Section 3. We first summarize the main descriptive patterns in the data—trend, seasonality, and the pandemic-related disruption—to motivate the modeling and uncertainty considerations. We then present benchmarking results for point forecasts under leakage-free designs (strict 2025 holdout and rolling-origin validation), followed by statistically defensible forecast comparisons. Finally, we report prediction-interval backtests, emphasizing empirical coverage, sharpness, and proper scoring, and discuss what these imply for deployment-oriented reliability and operational monitoring.

### *3.1. Trend and Seasonality*

Figure 2 shows Türkiye's monthly total aircraft movements (including overflights) over 2018–2025. The series exhibits strong annual seasonality with recurring within-year peaks and troughs, alongside a sharp disruption in 2020, followed by a multi-year recovery and a higher post-recovery level in 2023–2025. This pattern in Figure 2 indicates a setting where both the mean level and uncertainty may evolve over time, motivating regime-aware evaluation and caution in interpreting nominal model-based uncertainty without empirical checks.

Figure 3 summarizes the annual totals obtained by aggregating the reconstructed month-specific aircraft-movement series (including overflights) over 2018–2025. Two features are salient. First, the 2020 value shows a pronounced discontinuity relative to 2018–2019, indicating a system-wide contraction rather than a marginal fluctuation. Second, the subsequent years exhibit a sustained recovery path: totals increase monotonically from 2020 through 2025, with

the series surpassing pre-2020 levels by the end of the study window. Presenting annual aggregates serves both as a macro-level validation of the monthly reconstruction procedure and as an operational context for interpreting forecast errors, since planning-relevant deviations are ultimately realized in absolute annual volumes rather than only in relative monthly metrics.

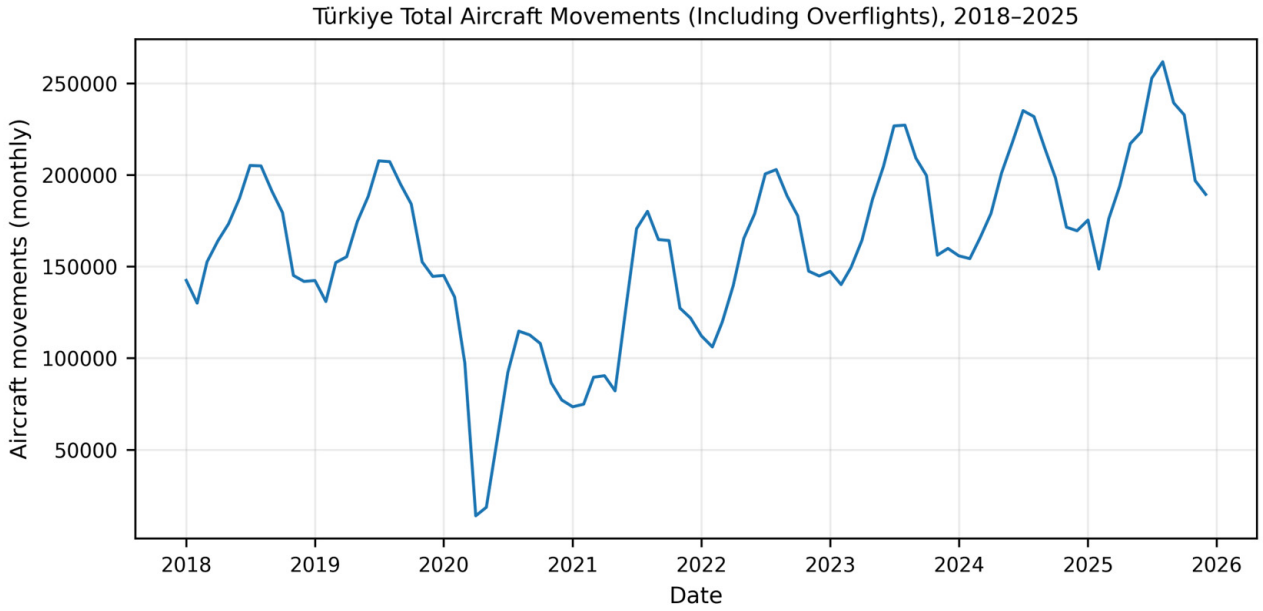


**Figure 2.** Monthly aircraft movements in Türkiye (total including overflights), 2018–2025.

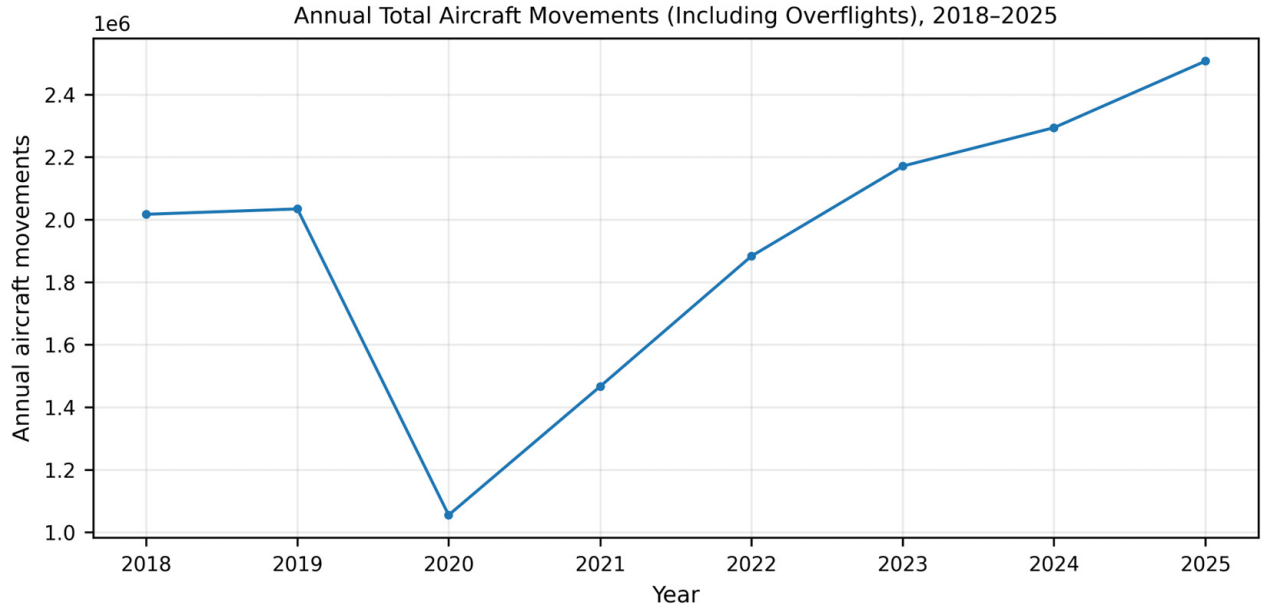


**Figure 3.** Annual totals derived from reconstructed monthly increments, 2018–2025.

### *3.2. Holdout Performance (2025)*

Holdout evaluation provides a deployment-style check because it tests models on a genuinely future period that was not used for fitting or model selection. We used the calendar year 2025 as a strict holdout to approximate an operational setting in which models are estimated on historical data and then asked to forecast the next unseen year. Table 1 summarizes the point-forecast performance using MAE, RMSE, MAPE, and sMAPE.

**Table 1.** 2025 holdout forecasting accuracy.

| Model | MAE | RMSE | MAPE | sMAPE |
|---|---|---|---|---|
| Seasonal naïve | 18,710 | 20,612 | 8.79% | 9.26% |
| ETS (Holt–Winters) | 28,704 | 32,182 | 13.36% | 14.51% |
| SARIMA | 6860 | 9474 | 3.63% | 3.52% |

Table 1 shows that SARIMA yields the lowest errors among the benchmarked models on the 2025 holdout across all four reported measures. The magnitude of the advantage is economically relevant, especially relative to the seasonal naïve benchmark, but the holdout contains only 12 monthly observations. Accordingly, the table is interpreted as strong descriptive evidence for better out-of-sample performance rather than as a standalone basis for broad inferential claims.

### *3.3. Rolling-Origin Cross-Validation (RO-CV)*

Rolling-origin cross-validation evaluates performance across multiple deployment-like origins by repeatedly expanding the training window and testing on the next unseen annual block. This design preserves temporal order and reduces the risk of information leakage. In our implementation, four yearly folds are used, which provide a useful but still limited view of temporal stability.

Table 2 summarizes the rolling-origin results across folds. Errors are highest in the earliest post-disruption fold, which is consistent with the transitional dynamics of the recovery period. In later folds, errors decline and become more stable, suggesting that predictability improved as seasonal patterns re-emerged. Because only four folds are available, these results should be read as directional evidence of temporal robustness rather than as an exhaustive stability assessment.

**Table 2.** RO-CV fold summary (selected SARIMA by AIC).

| Test Year | Selected Order | AIC | MAPE | sMAPE |
|---|---|---|---|---|
| 2022 | $(1,1,2) \times (0,1,1)_{12}$ | 459.03 | 11.65% | 11.61% |
| 2023 | $(2,1,1) \times (0,1,1)_{12}$ | 1229.05 | 3.64% | 3.69% |
| 2024 | $(2,1,2) \times (0,1,1)_{12}$ | 981.25 | 7.39% | 7.08% |
| 2025 | $(2,1,2) \times (0,1,1)_{12}$ | 1237.19 | 3.63% | 3.52% |

Across folds, the selected SARIMA orders vary over time, indicating that short-lag dynamics and residual structure are not fully constant across regimes. This variability is substantively informative: the benchmark is better understood as a transparent model-selection procedure within a constrained family than as one universally fixed parameterization (Table 3).

**Table 3.** Diebold–Mariano tests on 2025 holdout (SARIMA vs. baselines).

| Comparison | Loss | DM Statistic | *p*-Value |
|---|---|---|---|
| SARIMA vs. Seasonal Naïve | Squared error | −2.629 | 0.023 |
| SARIMA vs. ETS (Holt–Winters) | Squared error | −3.087 | 0.010 |
| SARIMA vs. Seasonal Naïve | Absolute error | −3.094 | 0.010 |
| SARIMA vs. ETS (Holt–Winters) | Absolute error | −3.998 | 0.002 |

### *3.4. Structural Break Evidence*

Table 4 reports evidence against full stability during the pandemic period. The Chow test in April 2020 is used here as a targeted diagnostic motivated by the timing of the

COVID-19 shock rather than as a general break-discovery procedure. The CUSUM results likewise indicate broader instability in parameter behavior over time. Taken together, these diagnostics are interpreted as supportive evidence of regime change rather than as definitive breakpoint identification.

**Table 4.** Structural break and stability tests around April 2020.

| Test | Statistic | *p*-Value | Note |
|---|---|---|---|
| Chow structural break | 14.446 | $4.96 \times 10^{-15}$ | OLS (trend + month dummies) stability |
| CUSUM | 2.212 | $1.13 \times 10^{-4}$ | OLS (trend + month dummies) residual stability |

### *3.5. Residual Diagnostics*

Residual diagnostics provide a compact check of whether the fitted SARIMA baseline leaves systematic structure unexplained and whether standard parametric uncertainty statements are likely to be credible. We evaluated three complementary properties of the one-step-ahead residuals. First, remaining serial dependence is assessed using the Ljung–Box Q test at seasonal lags, which is a standard lack-of-fit check in ARIMA-type models [8]. Second, we assessed distributional adequacy using the Jarque–Bera test, recognizing that nominal SARIMA prediction intervals typically rely on approximately Gaussian innovations. Third, we tested for time-varying volatility using Engle's ARCH-LM procedure since heteroskedasticity can cause nominal intervals to under-cover during high-variance periods even when the conditional mean is modeled reasonably well [10].

In our setting, the Ljung–Box results suggest no strong remaining autocorrelation at the examined seasonal lags, indicating that the mean dynamics are largely captured by the seasonal structure. However, evidence of non-normality (Jarque–Bera) and conditional heteroskedasticity (ARCH-LM) implies that purely model-based Gaussian intervals may be miscalibrated, particularly under regime shifts and recovery phases. Consequently, we treat prediction-interval quality as an empirical property and prioritize backtested coverage and sharpness when comparing uncertainty methods [23].

### *3.6. Prediction Interval Backtesting (2025)*

Operational decision-making in air traffic management requires not only accurate point forecasts but also uncertainty estimates that achieve their stated reliability in real time. We therefore evaluated prediction intervals on the strict 2025 holdout using a one-step-ahead backtesting design. At each month in 2025, intervals are generated from a model refit using information available up to the previous month, and performance is summarized over the 12 forecast instances. Table 5 reports the empirical coverage, average width, and interval score as complementary measures of interval quality.

**Table 5.** Prediction-interval backtesting results for the 2025 holdout.

| Method | Nominal Level | Empirical Coverage (2025) | Avg PI Width | Median PI Width | Interval Score |
|---|---|---|---|---|---|
| SARIMA nominal | 80% | 0.917 | 48,383 | 48,387 | 53,084 |
| Parametric bootstrap | 80% | 0.917 | 48,397 | 48,401 | 53,100 |
| Split conformal (cal 2023–2024) | 80% | 0.750 | 19,858 | 19,858 | 30,571 |
| Adaptive conformal (ACI; W = 24; $\gamma = 0.15$) | 80% | 0.833 | 35,379 | 25,124 | 45,147 |
| SARIMA nominal | 95% | 1.000 | 73,996 | 74,002 | 73,996 |
| Parametric bootstrap | 95% | 1.000 | 74,378 | 74,384 | 74,378 |
| Split conformal (cal 2023–2024) | 95% | 0.750 | 25,678 | 25,678 | 81,090 |
| Adaptive conformal (ACI; W = 24; $\gamma = 0.15$) | 95% | 0.917 | 65,760 | 69,272 | 91,320 |

The backtest results reveal a clear calibration-sharpness trade-off across methods. Nominal SARIMA intervals and parametric bootstrap intervals achieve relatively high empirical coverage, but they do so with comparatively wide bands. Split conformal

intervals are notably sharper, yet they under-cover on this holdout, which is consistent with the difficulty of maintaining nominal coverage under temporal dependence and regime shift. Adaptive conformal inference moves coverage closer to target levels at a moderate width cost, making it a practically useful compromise in this setting.

$$IS_\alpha(L_t, U_t; y_t) = (U_t - L_t) + \frac{2}{\alpha}(L_t - y_t)\mathbf{1}\{y_t < L_t\} + \frac{2}{\alpha}(y_t - U_t)\mathbf{1}\{y_t > U_t\} \quad (21)$$

To summarize prediction-interval performance with a single, interpretable criterion, we report the interval score (Winkler score), which is a proper scoring rule that jointly reflects sharpness and calibration. For each month $t$, let $[L_t, U_t]$ denote the $(1-\alpha)$ prediction interval and $y_t$ the realized observation. The score is computed as $IS_\alpha(L_t, U_t; y_t) = (U_t - L_t) + \frac{2}{\alpha}(L_t - y_t)1\{y_t < L_t\} + \frac{2}{\alpha}(y_t - U_t)1\{y_t > U_t\}$. The first term $(U_t - L_t)$ equals the interval width and captures sharpness, favoring narrower intervals when coverage is maintained. The indicator function $1\{\cdot\}$ equals 1 when its condition is true and 0 otherwise; if the realized value falls below the interval $(y_t < L_t)$, the score adds a penalty proportional to the miss distance $(L_t - y_t)$, and if it exceeds the interval $(y_t > U_t)$, it adds a penalty proportional to $(y_t - U_t)$. Both penalties are scaled by $2/\alpha$, so misses are penalized more strongly for higher-confidence intervals (smaller $\alpha$). In Table 5, we report the average interval score over the 12 one-step-ahead forecasts in the 2025 holdout for each nominal level (80% and 95%), where lower values indicate a more favorable balance between reliable coverage and informative (narrow) uncertainty bounds.

### *3.7. Three-Month Forecast*

To illustrate near-term planning use beyond the evaluation window, Table 6 reports the fixed-origin SARIMA projections for 2026 Q1 based on the final model estimated using all observations through December 2025. These forecasts are descriptive extensions of the benchmark rather than additional validation evidence. As expected, the interval width increases with the forecast horizon as uncertainty accumulates.

**Table 6.** Three-month forecast for 2026 Q1 with 80% and 95% prediction intervals.

| Month | Forecast | PI80_Low | PI80_High | PI95_Low | PI95_High |
|---|---|---|---|---|---|
| January 2026 | 191,636 | 168,051 | 215,222 | 155,566 | 227,707 |
| February 2026 | 179,741 | 145,205 | 214,277 | 126,923 | 232,559 |
| March 2026 | 198,181 | 155,578 | 240,785 | 133,025 | 263,338 |

## 4. Discussion

### *4.1. Main Findings*

The empirical results indicate that SARIMA provides the strongest overall point-forecast performance within the benchmark set considered in this study. Across both the strict 2025 holdout and the expanding-window rolling-origin evaluation, SARIMA outperforms the seasonal naïve and ETS alternatives. This finding is practically meaningful because both comparator models remain credible reference benchmarks for a seasonal monthly operational series.

At the same time, these results should be interpreted in proportion to the design. The 2025 holdout contains only 12 monthly observations, and the rolling-origin evaluation includes only four annual folds. For this reason, the present findings support the use of SARIMA as the strongest benchmark within the current framework, but they should not be overstated as universal evidence of superiority under all possible forecasting environments.

### *4.2. Interpreting the Forecasting Environment in a Regime-Aware Way*

A central implication of the results is that the 2018–2025 national air traffic series is unlikely to be well represented by one fully stable regime. The pandemic shock and the subsequent recovery phase introduce a forecasting environment in which level shifts, changes in seasonal structure, and possible parameter instability are all plausible. The Chow and CUSUM diagnostics are consistent with this interpretation, although they are used here as supportive diagnostic tools rather than as definitive break-identification procedures.

From an applied perspective, this matters because instability changes how forecast systems should be judged. In a regime-aware environment, forecast validation should not rely solely on retrospective in-sample adequacy. Instead, it should emphasize time-respecting out-of-sample evaluation, regular recalibration, and empirical checking of interval behavior under changing conditions.

### *4.3. Uncertainty as a First-Class Forecasting Output*

The interval results reinforce the idea that uncertainty should be treated as a first-class forecasting output rather than as a secondary byproduct of a fitted point-forecast model. In the present application, nominal SARIMA intervals and bootstrap intervals tend to be conservative, while split conformal intervals are sharper but under-cover. ACI provides a more balanced result, improving empirical coverage relative to split conformal without becoming as wide as the most conservative parametric alternatives.

These findings are useful, but they should also be interpreted carefully. The split conformal procedure is calibrated on only 24 monthly observations from 2023–2024, which limits the stability of the estimated conformal quantiles in the presence of dependence and distributional shift. Likewise, the ACI results are application-specific and should not be generalized as a universal ranking of uncertainty methods.

### *4.4. Contribution of the Study*

The main contribution is a reproducible forecasting baseline combined with leakage-free evaluation and interval backtesting. This contribution has four parts: an explicit reconstruction procedure for deriving monthly increments from cumulative official reporting; a leakage-free evaluation design based on a strict 2025 holdout and expanding-window rolling-origin validation; a regime-aware interpretation of the series; and interval backtesting under the same time-respecting framework.

### *4.5. Limitations*

Several limitations should be acknowledged clearly. First, the proposed framework is intentionally univariate and does not incorporate exogenous drivers that may influence national air traffic dynamics. While this design supports transparency, auditability, and reproducibility, it necessarily limits both the explanatory scope and potential predictive gains of the model. National air traffic demand is shaped by a range of external factors, including tourism flows, macroeconomic conditions, fuel prices, geopolitical developments, airspace restrictions, regulatory changes, and airport or network capacity constraints. Because these drivers are not explicitly modeled, the present results should be interpreted as establishing a transparent operational baseline rather than a fully explanatory forecasting system. Incorporating exogenous information through SARIMAX-type specifications, dynamic regression models, or hybrid frameworks represents a natural and important direction for future research, particularly in environments characterized by regime shifts and demand volatility.

Second, the rolling-origin validation includes only four annual folds. Although this structure remains leakage-free and operationally interpretable, it provides only a limited basis for assessing finer-grained temporal stability relative to denser monthly or quarterly origin designs. Third, the inferential role of the Diebold-Mariano tests is constrained by the short 2025 holdout horizon. Because the comparison is based on only 12 monthly observations, the DM results should be treated as supportive rather than definitive evidence.

Fourth, the split conformal calibration window is short, containing only 24 monthly observations. This small calibration set may yield unstable quantile estimates and limit the strength of any conclusion regarding conformal interval performance. Fifth, the benchmark set is intentionally centered on transparent statistical baselines and does not yet include machine learning alternatives or hybrid models. Likewise, the present study does not provide a formal comparison of computational efficiency, runtime, memory usage, update latency, or implementation complexity across model classes.

### *4.6. Future Extensions*

The most immediate extension of the present framework is the incorporation of exogenous drivers. Relevant candidates include tourism flows, exchange rates, fuel prices, inflation, GDP-related indicators, geopolitical disruption proxies, airspace policy changes, and capacity constraints. Such information could be introduced through SARIMAX-type specifications, dynamic regression frameworks, or hybrid pipelines that preserve the leakage-free evaluation principles established here.

A second extension concerns the benchmark set itself. Future work should compare transparent statistical baselines with broader machine learning and hybrid alternatives, including Prophet, gradient-boosting models, and sequence-based approaches, provided that all such models are evaluated under the same time-respecting validation design. A third extension concerns deployment-oriented benchmarking, including training time, inference latency, memory footprint, update cost, and implementation complexity.

### *4.7. Practical Implications*

For institutional forecasting practice, the main value of this study is not only that it identifies a relatively strong benchmark model, but that it demonstrates a more disciplined evidence standard for operational forecasting. The proposed framework provides a reproducible baseline with leakage-free evaluation and empirically validated uncertainty.

In this sense, the framework proposed here is best understood as a reproducible reference model for monthly planning and monitoring. Its role is to provide a credible baseline against which future multivariate, machine-learning, and deployment-oriented extensions can be compared.

## 5. Conclusions

This study develops a transparent and reproducible forecasting baseline for Türkiye's monthly national aircraft movements using public DHMI statistics for 2018–2025. Its main methodological contribution is the explicit reconstruction of month-specific increments from cumulative within-year reporting, combined with simple audit checks that make the data-preparation stage traceable and operationally interpretable.

Within the present benchmark design, SARIMA provides the strongest overall point-forecast performance relative to the seasonal naïve and ETS baselines under leakage-free evaluation. However, inferential claims based on the 2025 holdout should remain cautious because that comparison relies on only 12 monthly observations. For this reason, the Diebold–Mariano results are treated as supportive rather than definitive evidence, while

the broader benchmark pattern from the holdout and rolling-origin results provides the main basis for interpretation.

The structural-instability diagnostics further suggest that the series is not well characterized by a single stable regime over 2018–2025. These findings are best read as diagnostic evidence motivating a regime-aware forecasting posture rather than as exhaustive breakpoint identification. Under such conditions, uncertainty should be validated empirically rather than accepted as a nominal byproduct of a fitted model.

The interval backtests show that nominal and bootstrap intervals are conservative, split conformal intervals are sharper but under-cover, and adaptive conformal inference offers a more balanced result in this application. Because the calibration and evaluation windows are short, these findings should be interpreted as application-specific rather than as universal rankings of interval methods.

Overall, the proposed framework should be interpreted as a transparent benchmark for operational forecasting, not as a complete representation of the wider economic and policy determinants of national air traffic. Future work should extend this benchmark with exogenous drivers, machine learning and hybrid alternatives, and explicit deployment-oriented benchmarking of computational efficiency and operational costs under the same leakage-free evaluation principles used here.

**Author Contributions:** Conceptualization, R.K. and M.Ş.; methodology, M.K. and R.K.; software, R.K. and M.K.; validation, R.K., M.K. and M.Ş.; formal analysis, S.A.H.; investigation, M.Ş.; resources, S.A.H.; data curation, R.K.; writing—original draft preparation, R.K. and M.K.; writing—review and editing, M.Ş. and R.K.; visualization, M.Ş.; supervision, M.K.; project administration, S.A.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** All raw monthly statistics used in this study are publicly available from the DHMİ 'İstatistikler' portal (https://dhmi.gov.tr/Sayfalar/Istatistikler.aspx) under 'Türkiye Geneli Havalimanı İstatistikleri' (accessed on 14 February 2026). We accessed and downloaded the monthly 'Toplam Uçak' spreadsheets covering January 2018 to December 2025 on 2 February 2026. DHMİ notes that monthly releases may be cumulative within the calendar year; we reconstructed month-specific increments via within-year differencing, as described in Section 2.1. The processed month-specific dataset and the scripts used for ETL, model fitting, and evaluation are available from the corresponding author upon reasonable request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Hyndman, R.J.; Khandakar, Y. Automatic Time Series Forecasting: The forecast Package for R. *J. Stat. Softw.* **2008**, *27*, 1–22. [CrossRef]
2. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
3. Hopfe, D.H.; Lee, K.; Yu, C. Short-Term Forecasting Airport Passenger Flow during Periods of Volatility: Comparative Investigation of Time Series vs. Neural Network Models. *J. Air Transp. Manag.* **2024**, *115*, 102525. [CrossRef]
4. He, K.; Ji, L.; Wu, C.W.D.; Tso, K.F.G. Using SARIMA–CNN–LSTM Approach to Forecast Daily Tourism Demand. *J. Hosp. Tour. Manag.* **2021**, *49*, 25–33. [CrossRef]
5. Bergmeir, C.; Benítez, J.M. On the Use of Cross-Validation for Time Series Predictor Evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [CrossRef]
6. Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **1995**, *13*, 253–263. [CrossRef] [PubMed]
7. Chow, G.C. Tests of Equality between Sets of Coefficients in Two Linear Regressions. *Econometrica* **1960**, *28*, 591. [CrossRef]
8. Ljung, G.M.; Box, G.E.P. On a Measure of Lack of Fit in Time Series Models. *Biometrika* **1978**, *65*, 297–303. [CrossRef]
9. Jarque, C.M.; Bera, A.K. Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals. *Econ. Lett.* **1980**, *6*, 255–259. [CrossRef]

10. Engle, R.F. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica* **1982**, *50*, 987. [CrossRef]
11. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*, 2005th ed.; Springer: New York, NY, USA, 2005; ISBN 9780387001524.
12. Xu, C.; Xie, Y. Conformal Prediction Interval for Dynamic Time-Series. *Proc. Mach. Learn. Res.* **2020**, *139*, 11559–11569.
13. Gibbs, I.; Candès, E. Adaptive Conformal Inference under Distribution Shift. *arXiv* **2021**, arXiv:2106.00170. [CrossRef]
14. Brown, R.L.; Durbin, J.; Evans, J.M. Techniques for Testing the Constancy of Regression Relationships over Time. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1975**, *37*, 149–163. [CrossRef]
15. Box, G.E.P.; Jenkins, G.M.; Reinsel, G.C.; Ljung, G.M. *Time Series Analysis: Forecasting and Control*, 5th ed.; John Wiley & Sons: Nashville, TN, USA, 2015; ISBN 9781118675021.
16. Tashman, L.J. Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review. *Int. J. Forecast.* **2000**, *16*, 437–450. [CrossRef]
17. Hyndman, R.J.; Athanasopoulos, G. *Forecasting: Principles and Practice*, 3rd ed.; OTexts: Melbourne, Australia, 2021. Available online: https://otexts.com/fpp3/ (accessed on 10 January 2026).
18. Hyndman, R.J.; Koehler, A.B. Another Look at Measures of Forecast Accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [CrossRef]
19. Harvey, D.; Leybourne, S.; Newbold, P. Testing the Equality of Prediction Mean Squared Errors. *Int. J. Forecast.* **1997**, *13*, 281–291. [CrossRef]
20. Romano, Y.; Patterson, E.; Candès, E.J. Conformalized Quantile Regression. *arXiv* **2019**, arXiv:1905.03222.
21. Hamilton, J.D. *Time Series Analysis*; Princeton University Press: Princeton, NJ, USA, 1994; ISBN 9780691218632.
22. Bai, J.; Perron, P. Computation and Analysis of Multiple Structural Change Models. *J. Appl. Econ.* **2003**, *18*, 1–22. [CrossRef]
23. Gneiting, T.; Katzfuss, M. Probabilistic Forecasting. *Annu. Rev. Stat. Appl.* **2014**, *1*, 125–151. [CrossRef]
24. Taylor, S.J.; Letham, B. Forecasting at Scale. *Am. Stat.* **2018**, *72*, 37–45. [CrossRef]
25. Gneiting, T.; Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *J. Am. Stat. Assoc.* **2007**, *102*, 359–378. [CrossRef]
26. Barber, R.F.; Candès, E.J.; Ramdas, A.; Tibshirani, R.J. Predictive Inference with the Jackknife+. *Ann. Stat.* **2021**, *49*, 486–507. [CrossRef]
27. Angelopoulos, A.N.; Bates, S. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv* **2021**, arXiv:2107.07511.
28. Available online: https://www.eurocontrol.int/library (accessed on 14 February 2026).
29. Couto, R.A.; Maçaira Louro, P.M.; Cyrino Oliveira, F.L. Forecasting Wind Speed Using Climate Variables. *Forecasting* **2025**, *7*, 13. [CrossRef]
30. Ackermann, A.E.F.; Fani, V.; Bandinelli, R.; Sellitto, M.A. Short-Term Prediction in an Emergency Healthcare Unit: Comparison between ARIMA, ANN, and Logistic Map Models. *Forecasting* **2025**, *7*, 52. [CrossRef]
31. Soldatenko, S.; Alekseev, G.; Loginov, V.; Angudovich, Y.; Danilovich, I. Climate Indices as Potential Predictors in Empirical Long-Range Meteorological Forecasting Models. *Forecasting* **2026**, *8*, 9. [CrossRef]