Transport Economics and Management 4 (2026) 400–411

Contents lists available at ScienceDirect

Transport Economics and Management

journal homepage: www.journals.elsevier.com/transport-economics-and-management

Timing-aware container-throughput forecasting under information and uncertainty constraints

Mehdi Safaei *

Department of Logistics Management, Faculty of Economics, Administrative and Social Sciences, Istanbul Gelisim University, Istanbul, 34310, Türkiye

ARTICLE INFO

Keywords:
Container throughput forecasting
Information timing
Rolling-origin validation
Nested time-series tuning
Forecast comparison
Clark-West test
Sequential conformal prediction

ABSTRACT

Monthly container throughput in Türkiye is examined for one- and three-month forecast horizons using 172 observations from 2012M01 to 2026M04 and an expanding-window evaluation. For the national series, Ridge regularization is selected inside each outer training window, while component comparisons keep the Ridge penalty fixed across origins. Because the h= 1 forecast origin follows publication of the previous month's UAB statistics, it occurs after the target month has begun and is therefore interpreted operationally as a current-month nowcast rather than a forecast issued before the target month. Among the national specifications, the level-target operational-history Ridge gives the lowest forecast error. Compared with seasonal naive, mean MAE is lower by 31.2% at h= 1 and 27.0% at h= 3. These gains fall to about 20% after excluding February–March 2024, and the Harvey–Leybourne–Newbold-corrected tests do not reach the 5% level. Calendar-lagged trade does not lower national MAE. At h= 1, MAE is 1.9% higher even though the Clark–West test is significant ($p = 0.027$). At h= 3, MAE is 11.1% higher and $p = 0.845$. None of the 28 component-level lagged-trade additions is FDR-supported. In the retrospective same-period experiment, contemporaneous trade lowers loading-total MAE by 6.0% ($p = 0.004$; $q = 0.027$), but these values are unavailable at the forecast origin. For seasonal-log-difference targets, only Elastic Net improves on seasonal naive at both horizons, while all transformed-target models remain less accurate than the level-target Ridge. Sequential conformal prediction gives 82.1% coverage at h= 1 and 89.3% at h= 3 for the 80% nominal level.

1. Introduction

Monthly container demand has to be estimated before the actual situation of the next month is fully known by the port system. Staffing expectations, yard use, equipment deployment, workload, and service planning are all influenced by this early estimate. Türkiye has a regular national operational record for container activity. The Directorate General of Maritime Affairs (UAB) publishes monthly data on loading, unloading, export and import flows, cabotage, and transit [1]. These series are used as the operational history in this study. Trade statistics may also provide useful information because cargo movements are related to merchandise trade. But this relationship alone is not sufficient to make a variable useful for forecasting. A predictor can contribute to the forecast only if its value is already available when the forecast is made.

Port-throughput research has pursued richer algorithms and richer predictor sets. Neural networks, recurrent architectures, hybrid models, forecast combinations, and explainable machine-learning methods are represented in recent work [2–8]. Other studies move beyond past throughput. Eskafi et al. [9] connect port activity with macroeconomic conditions and identify national export trade as an important influence on containerized throughput; Morales-Ramírez, Gracia and Mar-Ortiz [10] analyze national and coastal cargo throughput with macroeconomic indicators; Zeng and Xu [11] use hinterland and foreland information in a deep-learning model. These studies make external predictors worth testing. They also expose a practical complication: month t on a statistical series is not necessarily information that existed when a month-t forecast had to be issued.

That complication is common in economic data. A statistic refers to one period, appears later, and may then be revised.

Therefore, calendar alignment can produce a completely regular dataset while still containing information that was not available at the forecast origin. In out-of-sample forecasting research, this problem is known as leakage [12–14]. In nowcasting, the timing issue is seen more directly. In Hopp [15], econometric and machine-learning methods are compared when some important releases are published with delay.

* Correspondence to: Cihangir, Petrol Ofisi Cd. No: 3, 34310 Avcılar, Istanbul, Türkiye.
E-mail address: msafaei@gelisim.edu.tr.

https://doi.org/10.1016/j.team.2026.10.005
Received 12 August 2026; Received in revised form 19 October 2026; Accepted 19 October 2026
Available online 20 October 2026
2949-8996/© 2026 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Chinn, Meunier and Stumpner [16] also estimate world trade using real-time data and indicators available earlier, before the official volume series is published. The same issue is also seen in the Türkiye data in a port setting. Recent UAB releases are published around 10–15 days after the end of the month, while the trade series used in this study become available later (Section 2). Therefore, simply matching observations by reference month cannot show that the trade information was available to the forecaster at the time of the forecast. A short monthly sample raises two further concerns. Model ranking is one. Selection made after seeing an evaluation period is different from selection confined to the training data, and rankings can also depend on how the target is represented. Munim et al. [17] report exactly this sort of contrast in a port application: Prophet is strongest in sample, whereas forecast combinations perform better out of sample. Forecast uncertainty is the other concern. A point forecast by itself gives no range for planning. Conformal prediction offers distribution-free intervals [18,19], but a time-series application cannot use a calibration error before its target is observed. Related transport work also separates forecast accuracy from operational trigger value [20], while De Santis, Giagnorio and Matteucci [21] show that uncertainty receives uneven treatment in transport appraisal.

The empirical problem is therefore more specific than simply selecting a forecasting algorithm. The main issue is to evaluate national and component throughput using an information set that follows actual publication timing, and then examine whether trade variables add useful information beyond operational history and how much uncertainty remains in the forecasts. The sample contains 172 monthly observations from 2012M01 to 2026M04. A common period from 2024M01 to 2026M04 gives 28 months for comparison. Forecast origins are defined according to UAB publication timing, and the rolling-origin design uses only information that was available at each relevant origin. Model selection, statistical comparison, and uncertainty calibration are explained in Section 3.

The analysis considers three research questions. RQ1 examines how accurately operational history forecasts national throughput and the seven operational flows under rolling-origin evaluation. RQ2 asks whether calendar-lagged trade provides additional forecasting information after operational history is already included. RQ3 considers the uncertainty that remains when conformal calibration follows the same chronological availability rule.

The main empirical results show a fairly consistent pattern. Among the national specifications, the operational-history Ridge estimated on the level target gives the lowest forecast error. Its mean MAE is 31.2% lower than seasonal naive at h= 1 and 27.0% lower at h= 3. A small number of months, however, account for an important part of this mean difference, and the Harvey–Leybourne–Newbold-corrected tests do not reach the 5% significance level. The calendar-lagged trade variables do not show a consistent forecasting benefit. For the national series, MAE is higher at both forecast horizons. At h= 1, the Clark–West test is significant ($p = 0.027$), but this result is not replicated at h= 3. After Benjamini–Hochberg adjustment, none of the 28 component-level trade additions is FDR-supported. Contemporaneous trade gives a different result: in the retrospective same-period experiment, loading-total MAE decreases by 6.0% (BH $q = 0.027$). These trade observations become available only after both the forecast origin and the corresponding throughput release. Sequential same-horizon calibration is then used to express the remaining uncertainty as throughput-uncertainty margins in TEU, with binomial-reference confidence intervals reported for the observed coverage proportions.

Section 1.1 develops the forecasting and transport context for these questions. Section 2 documents the data and the dates at which they become usable. Section 3 describes the forecasting and evaluation design. Empirical results appear in Section 4, followed by robustness and limitations in Section 5, operational implications in Section 6, and the conclusion in Section 7.

1.1. Research context and positioning

Container-throughput forecasting has moved well beyond simple extrapolation, but greater model flexibility has not produced a single preferred approach. Milenković et al. [3] combine neural networks with metaheuristic optimization, Tan et al. [5] use recurrent neural networks, and Shankar, Punia and Ilavarasan [4] develop a deep-learning formulation. Other studies compare machine-learning and conventional methods [8], construct forecast combinations [7], or compare statistical, machine-learning, and explainable AI models [2]. When different forecasting methods are compared, evaluation design also matters. Munim et al. [17] used monthly data from Shanghai, Busan, and Nagoya. In their results, Prophet performed better in sample, while combined models performed better out of sample. So a model may work well during the estimation period but not keep the same advantage when later observations are forecast.

Forecasting may also improve when information outside throughput history is used. Wang et al. [6] included hinterland economic indicators, port-transport variables, and transport-capacity factors in a multivariable hybrid system. Their study reported point forecasts and also uncertainty ranges. Zeng and Xu [11] used information from both the hinterland and foreland in a deep-learning model. Eskafi et al. [9] examined an Icelandic port and studied containerized throughput together with national export trade volume. Their mutual-information analysis identified national export trade as the main macroeconomic influence. In Mexico, Morales-Ramírez, Gracia and Mar-Ortiz [10] used VAR and SARIMA models for monthly port cargo at national and coastal levels. They also examined its relationship with macroeconomic indicators. These studies give a reason to test information beyond throughput history. But for operational forecasting, matching data by reference month is not enough. The information must also be available when the forecast is actually made.

In real-time forecasting, publication date is an important issue. An observation may belong to the correct reference month but still not be available when the forecast is produced. Tashman [14] considers genuine out-of-sample testing as a major part of forecast evaluation. Bergmeir and Benítez [13] and Hewamalage, Ackermann and Bergmeir [12] also discuss validation and leakage problems in time-series prediction. The same problem of data arrival appears directly in macroeconomic nowcasting. Hopp [15] evaluates 12 econometric and machine-learning methods for US GDP in a setting where timeliness is part of the forecasting problem. Chinn, Meunier and Stumpner [16] use real-time datasets and indicators available earlier to estimate world trade before the delayed official trade-volume measure is released. In their application, regression-based machine-learning variants perform better than common tree-based alternatives. For the port problem considered in this study, the point is straightforward. Predictive usefulness needs to be evaluated according to the data vintage and release date that were actually available at the forecast origin.

How forecasts are compared can also change the conclusion. Model selection may appear stronger than it really is when the same evaluation period affects both model choice and final assessment. Rankings can also change when a different target representation is used. Munim et al. [17] provide a port example of the first issue. Hyndman and Koehler [22] introduced MASE for scale-free forecast accuracy. For inference on predictive accuracy, Diebold and Mariano [23] and Harvey, Leybourne and Newbold [24] provide out-of-sample comparison tests, while Clark and West [25] deal with comparisons involving nested models. When several related additions are tested, Benjamini and Hochberg [26] provide a procedure for controlling the false discovery rate. These methods address different questions, including forecast accuracy, model comparison, incremental value in nested models, and multiplicity. They therefore should not be treated as one common significance rule.

Container accounting provides another useful check. National throughput is observed directly, but the same total can also be obtained from its operating components. Hierarchical forecasting methods are

designed to reconcile forecasts across this type of structure [27,28], and Sanguri et al. [29] use coherent forecasting in port operations. In this study, the accounting identities are used for a more limited purpose. The direct national forecast is compared with the sum of the loading and unloading forecasts. An equal-weight average of these two forecasts is also reported as a simple combination. This is only an internal consistency comparison, and no formal reconciliation estimator is used.

Uncertainty around point forecasts is also considered in the methodological framework. Eskafi et al. [9] examined epistemic uncertainty when identifying variables related to port throughput. Wang et al. [6] reported uncertainty ranges together with their point forecasts. Conformal prediction treats this issue differently. The prediction intervals are based on empirical forecast errors and do not require a parametric assumption about the distribution of those errors [18,19]. Chronology changes how that idea can be used: a residual cannot enter calibration until its target exists. The short evaluation period creates uncertainty about the observed coverage proportion as well. This uncertainty is summarized using Clopper–Pearson intervals as a binomial reference; because rolling coverage indicators may be serially dependent, these intervals are not interpreted as exact confidence intervals under the time-series sampling process.

These literatures meet at the distinction tested here. An external economic indicator may contain information about port activity but arrive too late to improve the operational forecast. The Türkiye application therefore evaluates trade according to actual release timing, measures its incremental contribution against an operational-history baseline, and reports the uncertainty left around the forecasts. Contemporaneous trade contains a current-month signal in the retrospective analysis, yet the observed publication schedule prevents that signal from entering the forecast at the defined origin. The same issue arises whenever operational data are combined with externally published indicators whose reference dates precede their release dates.

2. Data and decision-time availability

The target is Türkiye's official monthly total container handling published by the Directorate General of Maritime Affairs (UAB). Recent entries in the UAB archive show April 2026 statistics released on 11 May 2026, May on 15 June, June on 13 July, July on 10 August, and August on 10 September. For the 2026 releases, the publication lag was around 10–15 days after the end of each month [1]. The timing for earlier years was not reconstructed and may have been different. Therefore, this recent pattern is used in the backtest as the operational timing convention. The forecast origin, or (o), is defined as the date when UAB publishes the statistics for month (o). For h= 1, the target is month o + 1, but the forecast origin is the UAB publication date for month o, which occurs approximately 10–15 days after month-end in the observed 2026 releases. Consequently, the h= 1 prediction is issued after month o + 1 has already begun. Operationally, it should therefore be interpreted as a current-month nowcast rather than as a forecast available before the target month starts. By contrast, h= 3 remains a forward multi-month forecast because its target month lies beyond the current calendar month at the forecast origin. Export and import data are taken from the OECD-derived FRED/ALFRED series XTEXVA01TRM667S and XTIMVA01TRM667S. In FRED, both series are described as monthly, seasonally adjusted, exchange-rate-converted US-dollar series. In the archived export view dated 15 May 2026, data are available through March 2026. In the 15 June update, the April 2026 value is added. Therefore, in the examination conducted in this study, June 15 is the first visible ALFRED vintage that shows the April export value. For imports, the April 2026 value appears on the public page in the 15 June update. However, an earlier import snapshot was not independently recovered. Therefore, this date is treated only as an update date [30,31]. Supplementary Table S3 also reports the observed publication dates in 2026 and their corresponding month-end lags.

Seven monthly operating flows sit beneath the national total: export loading, import unloading, loading total, unloading total, international total, cabotage, and transit. Two monthly identities are verified. National throughput equals loading plus unloading; it also equals international plus cabotage plus transit. They are used for the data audit and for the direct-versus-bottom-up consistency comparison.

Table 1 assigns each source an analytical role according to its observed release pattern. Lags of the UAB throughput series form the operational-history inputs, and the origin is anchored to the UAB release date for the latest observed month. Calendar-lagged trade specifications are treated as pseudo-real-time because the exact point-in-time vintage panel is incomplete and the trade series are revised and seasonally adjusted. Same-reference-month trade enters only the retrospective experiment. With this definition of origin, lag-1 trade is released too late; lag-2 is borderline because its release occurs around the same time as the UAB-based origin. The results of these data-quality and accounting-identity checks are summarized in Table 2.

3. Forecasting and evaluation design

Four parts of the design are kept separate: operational-history forecasting, calendar-lagged trade evaluation, model-family checks under an alternative target representation, and sequential uncertainty calibration. The national Ridge reference selects its regularization parameter by chronological inner validation within each outer training window. Component forecasts use Ridge with alpha= 10 fixed at every outer origin, and the same fixed specification is used when component trade predictors are added. Separate model-family tournaments are used as robustness checks.

3.1. Decision-time information sets

Let o denote the forecast origin, defined as the UAB release date for month o, and let $y_{(o+h)}$denote the h-step target. The horizon index is defined in monthly target steps relative to the latest observed UAB month. Under the observed publication schedule, h= 1 therefore corresponds operationally to current-month nowcasting, whereas h= 3 remains a genuinely forward forecast horizon. Information sets are defined relative to o:

$$\mathscr{A}_o = H_o(y) \tag{1}$$

$$\mathscr{B}_o^{cal} = T_o^{cal} \tag{2}$$

$$\mathscr{C}_o = \mathscr{A}_o \cup \mathscr{B}_o^{cal} \tag{3}$$

$$\mathscr{D}_o = \mathscr{A}_o \cup \{x_{o+1}^{\exp}, \; x_{o+1}^{imp}\} \tag{4}$$

Here, $H_o(y)$ collects operational-history features constructed only from throughput observable by origin o. T_o^{cal} collects the final-vintage calendar-lag trade features used in the sensitivity analysis. The predefined trade feature set includes lags 1, 2, 3, 6, and 12 months together with derived rolling and growth terms. Because lag-1 is released after o and lag-2 is borderline under the observed schedule, the calendar-lag exercise is more permissive to trade than an operational forecast. Trade nevertheless does not produce a reliable MAE gain. In Table 6, "system" denotes national-total UAB context features—lagged total TEU and derived rolling/year-on-year summaries—combined with component history and trade variables. $\mathscr{D}_o$ adds same-reference-month trade directly to operational history, not to $\mathscr{C}_o$, and is used only in the h= 1 retrospective same-period experiment reported in Supplementary Table S5.

3.2. Rolling-origin and nested time-series validation

Accuracy is evaluated by expanding-window rolling origins [14]. At an outer origin, training contains only targets already observable.

Table 1
Data sources, release-time status, and admissible analytical role.

Variable	Source	Frequency	Period	Decision-time / release status	Strict forecast role	Same-period role	Risk if misused
Total and seven TEU series	UAB	Monthly	2012M01–2026M04	Apr 11 May; May 15 Jun; Jun 13 Jul; Jul 10 Aug; Aug 10 Sep 2026 (~10–15 days after month-end). Origin = UAB release date.	Yes, after month-o release; h= 1 targets reference month o + 1.	Target only	Month-o TEU is not assumed available at month-end.
Exports (USD, exchange-rate converted; seasonally adjusted)	OECD series via FRED/ ALFRED	Monthly	2012M01–2026M04	Archived 15-May-2026 export view ends Mar-2026; 15-Jun update contains Apr-2026. Seasonally adjusted; revised.	Not assumed real-time feasible; calendar-lag sensitivity only	Retrospective same-period experiment only	Seasonal adjustment and final-vintage revisions can look ahead; reference month ≠ release date
Imports (USD, exchange-rate converted; seasonally adjusted)	OECD series via FRED/ ALFRED	Monthly	2012M01–2026M04	Apr-2026 appears in 15-Jun update; exact first import vintage not independently reconstructed. Seasonally adjusted; revised.	Not assumed real-time feasible; calendar-lag sensitivity only	Retrospective same-period experiment only	Seasonal adjustment and final-vintage revisions can look ahead; reference month ≠ release date
Container-type breakdown	UAB	Annual / cumulative	2012–2025	Not a monthly panel	No monthly modeling role	No	Do not infer monthly hierarchy
Port-authority breakdown	UAB	Annual / cumulative	2012–2025	Not a monthly panel	No monthly modeling role	No	Do not infer monthly hierarchy

Table 2
Data-quality and accounting-identity checks.

Check	Result
National monthly TEU observations	172; zero missing; zero duplicates
Export / import observations	172 each; zero missing
Total = Loading + Unloading, maximum absolute error	1.00 TEU
Total = International + Cabotage + Transit, maximum absolute error	1.00 TEU
Annual national total vs. container-type annual total, maximum absolute error	0.75 TEU
Annual national total vs. port-authority annual total, maximum absolute error	0.75 TEU
May 2026	Excluded, export-series alignment ends April 2026
Aggregation-consistency feasibility	Both accounting identities valid at monthly resolution; direct-national versus bottom-up consistency check is feasible and reported in Section 4.5

National models and the model-family robustness tournament select hyperparameters through chronological expanding-window validation inside that outer training sample. Component reference forecasts and component trade-addition tests do not retune the penalty: Ridge alpha= 10 remains fixed across origins, and no component hyperparameter is selected from the 28-month outer evaluation window. This keeps penalty selection from changing at the same time as the added predictor set.

$$\hat{y}_{o+h|o} = f_h(\mathscr{I}_{o,h}), \mathscr{I}_{o,h} \in \{\mathscr{A}_o, \quad \mathscr{C}_o\}, \quad h \in \{1,3\}; \textit{retrospective}h = 1 \\ : \quad \mathscr{I}_{o,1} = \mathscr{D}_o \tag{5}$$

The outer comparison period is 2024M01–2026M04, giving 28 target months at h= 1 and h= 3. At h= 3, forecast-comparison variance estimates use the serial-correlation correction appropriate to the horizon. Repeated use of this comparison period is treated as a limitation in Section 5.

3.3. Model families and tuning

The broader tournament includes seasonal naive, Ridge, Elastic Net, Random Forest, Extra Trees, histogram-based gradient boosting [32,33], and radial-basis-function SVR. Predictor preprocessing is fitted within each training window. The main tournament models log throughput levels. A separate diagnostic robustness exercise uses the common seasonal log difference $z_t = \log y_t - \log y_{t-12}$ for every model family. To keep predictors commensurate with that target, the transformed exercise uses lagged seasonal log differences and their rolling moments rather than level predictors, together with calendar terms known at the forecast origin. Forecasts are returned to TEU using $\hat{y}_t = y_{t-12}\exp(\hat{z}_t)$; for h= 3 the reconstruction anchor $y_{t-12} = y_{o-9}$ is already observed at origin o. SVR targets are standardized within each training window and inverse-transformed before level reconstruction. This transformed-target exercise is treated as a diagnostic robustness check, not as a replacement for the level-target reference forecast. The model families, prespecified tuning grids, and nested validation design are summarized in Table 3.

Same-period trade values are not part of the operational forecasts. Their ALFRED histories show release after the forecast origin; they enter only the retrospective information experiment.

3.4. Accuracy metrics

Let $e_i = y_i - \hat{y}_i$ denote the forecast error for test observation i. MAE, RMSE, MAPE, and sMAPE are reported on the original TEU scale as defined in Eqs. (6)–(7).

$$\text{MAE} = 1/n\sum_i |e_i|, \text{RMSE} = \sqrt{1/n\sum_i e_i^2} \tag{6}$$

$$\text{MAPE} = 100/n\sum_i |e_i/y_i|, \text{sMAPE} = 100/n\sum_i \frac{2|e_i|}{|y_i| + |\hat{y}_i|} \tag{7}$$

Mean absolute scaled error (MASE; [22]) provides the scale-free accuracy measure. In Eq. (8), n_o is the number of training observations available at origin o, the seasonal summation extends from s = 13 to n_o, and O denotes the outer test origins for the relevant horizon.

$$d_o = \frac{1}{n_o - 12}\sum_{s=13}^{n_o} |y_s - y_{s-12}|, \text{MASE} = \frac{1}{|O|}\sum_{o\in O}\frac{|y_{o+h} - \hat{y}_{o+h|o}|}{d_o}. \tag{8}$$

The same origin-specific MASE definition is used in every reported comparison. The error at an origin is scaled by the mean absolute seasonal-naive error computed only from training observations available there, after which the scaled errors are averaged across origins. Consequently, seasonal naive has test-period MASE of 1.106 at h= 1 and 1.108 at h= 3; the value 1 is the in-sample scaling reference, not a forced test-period result. To make the benchmark comparison transparent, the main tables also give MAE in TEU, seasonal-naive MAE over the same 28 test months, the percentage MAE difference, and a paired forecast-

Table 3
Model families, prespecified tuning grids, and nested validation design.

Model	Predictors	Prespecified search grid	Inner validation	Outer validation	Preprocessing / target
Seasonal naive	Lag-12 level	None	—	Expanding-window	Original TEU level
Ridge	Common lag/rolling/calendar features	alpha grid	Expanding-window	Expanding-window	X standardized; log-level primary target
Elastic Net	Same	alpha × l1_ratio grid	Expanding-window	Expanding-window	X standardized; log-level primary target
Random Forest	Same	depth/leaf/feature grid	Expanding-window	Expanding-window	Common features; log-level primary target
Extra Trees	Same	depth/leaf/feature grid	Expanding-window	Expanding-window	Common features; log-level primary target
Hist. Gradient Boosting	Same	depth/leaf/lr/iterations grid	Expanding-window	Expanding-window	Common features; Friedman [32]; sklearn implementation
RBF-SVR	Same	C × epsilon × gamma grid	Expanding-window	Expanding-window	X standardized; y standardized in seasonal-difference robustness

Note: Component-level Ridge forecasts in Tables 5 and 6 and Supplementary Table S5 use alpha = 10 at every origin. The alpha grid shown here applies only to the national models and the model-family robustness tournament.

comparison test based on the test-window errors.

3.5. Incremental information value and statistical inference

For each component–horizon pair, MAE_{base} refers to the operational-history Ridge and MAE_{aug} to the corresponding specification after the stated trade/system predictors are added. Eq. (9) defines Forecast Value Gain (FVG).

$$FVG = 100 \times \frac{MAE_{base} - MAE_{aug}}{MAE_{base}}, \tag{9}$$

FVG expresses the MAE reduction from the operational-history baseline to its augmented counterpart; a positive value means lower error. At national level, the trade tests compare nested Ridge specifications against the same baseline. Component tests use the fixed Ridge specification in both members of each pair, leaving the added predictors as the nested difference.

3.5.1. Paired tests and multiplicity control

Nested additions need a forecast-comparison statistic that accounts for the larger model containing the smaller one under the null. For month t, Eq. (10) forms the Clark–West adjusted loss differential from the baseline error e_1, the augmented-model error e_2, and forecasts $\hat{y}_1$ and $\hat{y}_2$. The one-sided test asks whether the added information has positive incremental predictive content. The 28 prespecified component comparisons—seven flows crossed with four additions—are treated as one family, and Benjamini–Hochberg [26] FDR adjustment is applied to the Clark–West p-values. Wilcoxon, bootstrap, conventional paired-loss, and dependence-aware maxT calculations remain robustness diagnostics. Supplementary Table S4 reports the Clark-West, BH, paired-loss, Wilcoxon, bootstrap, and Westfall-Young results for all 28 comparisons.

$$f_t^{CW} = e_{1,t}^2 - \left[e_{2,t}^2 - \left(\hat{y}_{1,t} - \hat{y}_{2,t}\right)^2\right], H_0 : E(f_t^{CW}) \le 0 \tag{10}$$

Non-nested comparisons, including the operational-history Ridge forecast against seasonal naive, use the Diebold–Mariano loss-differential framework [23]. Variance is estimated with Newey–West HAC and the Harvey–Leybourne–Newbold small-sample correction [24, 34]. The HAC truncation is h−1, which allows for serial correlation from overlapping multi-step errors at h= 3. Each comparison is interpreted using its designated primary test; the other tests are reported as robustness evidence. The two national trade-addition tests lie outside the 28-test component family and are interpreted jointly with the national MAE results in Section 4.2.

3.6. Direct national and bottom-up forecast comparison

Because national throughput equals loading plus unloading, h= 1 can be forecast through two independently estimated routes. The direct route models total throughput itself. The bottom-up route adds separate forecasts for loading total and unloading total. Their difference is examined as an aggregation-consistency diagnostic.

$$\hat{y}_t^{BU} = \hat{y}_t^{Load} + \hat{y}_t^{Unload} \tag{11}$$

An equal-weight average of the direct and bottom-up forecasts is reported as an additional combination. It is not MinT, WLS, OLS, or another projection-based hierarchical reconciliation estimator; those methods are outside this comparison.

$$\hat{y}_t^{COMB} = 0.5\hat{y}_t^{DIR} + 0.5\hat{y}_t^{BU}, g_t = \hat{y}_t^{DIR} - \hat{y}_t^{BU} \tag{12}$$

Paired absolute errors from the direct and bottom-up routes are compared with the Wilcoxon signed-rank test. The equal-weight combination is reported alongside them to show its position relative to both routes.

3.7. Sequential conformal uncertainty intervals

Conformal intervals are built separately for each horizon around the same TEU-only Ridge point forecast reported in Table 4; its Ridge penalty is chosen by inner expanding-window validation at each origin. Pre-2024 rolling-origin forecasts supply warm-up residuals before the test period begins. At origin o, the calibration pool for horizon h contains

Table 4
National forecast accuracy under matched nested Ridge specifications and the seasonal-naive benchmark.

h	Specification	MAE (TEU)	Unified MASE	S-naive MAE	MAE gain vs S-naive	Primary test
1	Seasonal naive	74,111	1.106	74,111	—	Reference
1	TEU-only Ridge	50,977	0.758	74,111	31.2%	HLN p = 0.087
1	Combined Ridge (calendar-lag sensitivity)	51,922	0.773	74,111	29.9%	CW vs TEU p = 0.027
3	Seasonal naive	74,111	1.108	74,111	—	Reference
3	TEU-only Ridge	54,110	0.806	74,111	27.0%	HLN p = 0.099
3	Combined Ridge (calendar-lag sensitivity)	60,133	0.896	74,111	18.9%	CW vs TEU p = 0.845

only absolute residuals from h-step forecasts whose target months are already observable; h= 3 residuals therefore join the pool later than h= 1 residuals. With nominal coverage 1−α and n calibration residuals, the finite-sample rank is k = ceil[(n + 1)(1 −α)], capped at n. Eq. (13) gives the chronological set, its order statistic, and the interval. Once a target becomes observable, its residual becomes available for subsequent origins.

$$A_{o,h} = \{r_{j,h} : t_j \le o\}, n_{o,h} = |A_{o,h}|, k = \min\{n_{o,h}, \text{ceil}[(n_{o,h}+1)(1-\alpha)]\},$$

$$q_{o,h} = r_{(k),h}, C_{o,h} = [\hat{y}_{o+h|o} - q_{o,h},\ \hat{y}_{o+h|o} + q_{o,h}] \quad (13)$$

Empirical coverage is reported together with two-sided 95% Clopper–Pearson intervals used as a binomial reference. These intervals would have their conventional exact binomial interpretation if the coverage indicators were independent Bernoulli observations with a common coverage probability. Because rolling-origin coverage indicators may be serially dependent, particularly at h= 3, the reported intervals are treated as descriptive binomial-reference intervals rather than exact confidence intervals for the dependent forecast sequence [35]. Interval half-width is expressed in TEU and termed the throughput-uncertainty margin.

3.8. Robustness checks

Robustness checks address four issues: target representation, release-time interpretation, dependence-aware multiplicity adjustment, and MASE scaling. In the seasonal-log-difference exercise, predictors are transformed consistently with the differenced target and the return to TEU levels is verified at both horizons. Its performance is assessed against seasonal naive and against the more accurate level-target Ridge before any conclusion is drawn about model complexity.

4. Results

The empirical results follow the sequence of the analysis: data timing, national accuracy, target representation, added trade information, the same-period experiment, direct-versus-bottom-up consistency, and sequential uncertainty.

4.1. Data audit and timing feasibility

The data audit finds all 172 monthly observations aligned, with both accounting identities closing to rounding tolerance. Publication timing is less tidy. The export archive dated 15 May 2026 ends in March 2026; the update dated 15 June contains April 2026. The import page also associates April 2026 with the 15 June update, although an earlier import-vintage snapshot was not independently recovered. Given these histories [30,31], trade-augmented results are interpreted as pseudo-real-time sensitivity evidence. For the operational-history forecasts, only UAB information already available at each forecast origin is used [1].

4.2. National forecasting performance

At h= 1, the operational-history Ridge in Table 4 records MAE = 50,977 TEU and unified MASE = 0.758. At h= 3, MAE = 54,110 TEU and unified MASE = 0.806. Seasonal naive is less accurate on mean absolute error: MAE = 74,111 TEU at both horizons, with origin-scaled test MASE of 1.106 at h= 1 and 1.108 at h= 3. The corresponding mean MAE reductions for Ridge are 31.2% and 27.0%. The paired HLN-corrected loss-differential tests give p = 0.087 at h= 1 and p = 0.099 at h= 3. Figs. 1–6 display the month-by-month absolute errors that underlie these aggregate values.

Trade augmentation does not improve national MAE. Combined Ridge reaches 51,922 TEU at h= 1, 1.9% worse than TEU-only Ridge, and 60,133 TEU at h= 3, 11.1% worse. The one-sided Clark–West adjusted-MSPE p-values are p = 0.027 at h= 1 and p = 0.845 at h= 3. At h= 1, the Clark–West test rejects in favor of lower adjusted squared-error loss for the model with trade. MAE is nevertheless higher for that model, and the same Clark–West result is not replicated at h= 3. The h= 1 rejection alone is therefore not treated as reliable national evidence that trade improves forecast accuracy. The seasonal-naive comparison is also influenced by one unusually large month. In February 2024 its error is 326,617 TEU because the rule reuses the low February 2023 level of 840,593 TEU against actual February 2024 throughput of 1167,210 TEU. February 2023 was also the period of the 6 February earthquakes. At Iskenderun Port, fire damage, damaged containers and later rehabilitation works were reported during this period [36]. Here, this timing is only noted as a coincidence. It is not taken as evidence that these events caused the national outlier. When February and March 2024 are removed from the comparison, the advantage of Ridge becomes smaller, reaching 20.1% for h= 1 and 20.3% for h= 3. Using median absolute error instead gives more limited differences, 5.5% and 13.1%, respectively. Supplementary Table S2 contains these

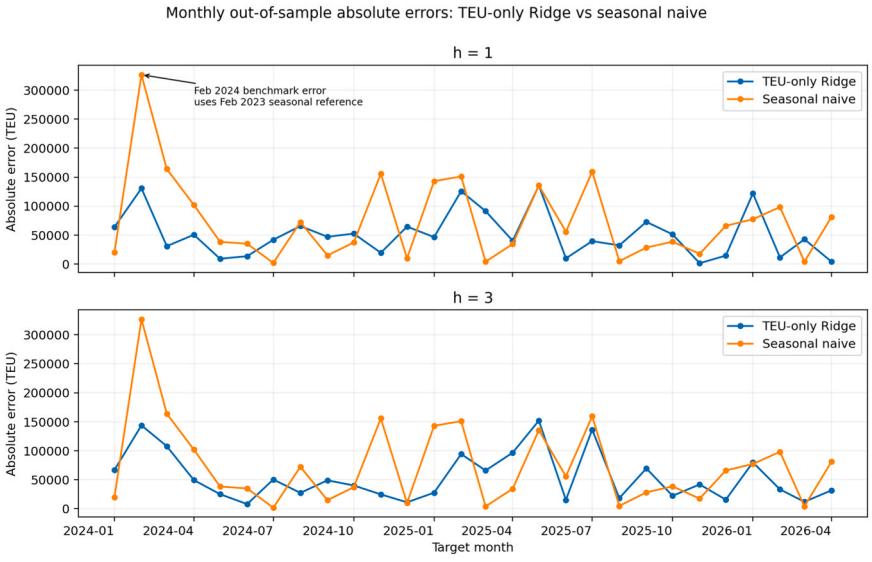


Fig. 1. Monthly out-of-sample absolute errors for the TEU-only Ridge reference and seasonal-naive benchmark at h= 1 and h= 3. The figure adds temporal error structure to the aggregate accuracy statistics in Table 4.

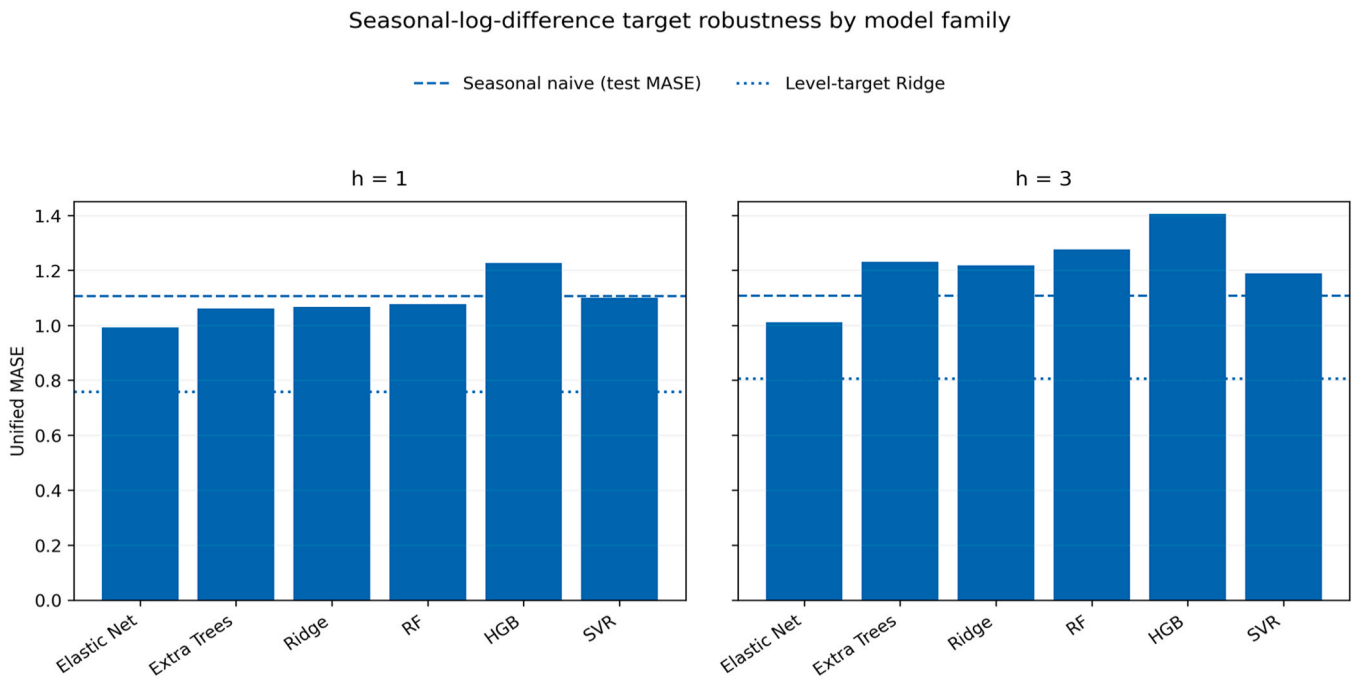


Fig. 2. Seasonal-log-difference target robustness by model family. Dashed lines show the actual test-period seasonal-naive MASE under the origin-specific scaling (1.106 at h=1; 1.108 at h=3), and dotted lines show the level-target Ridge reference. MAE-based benchmark comparisons are reported in Supplementary Table S1.

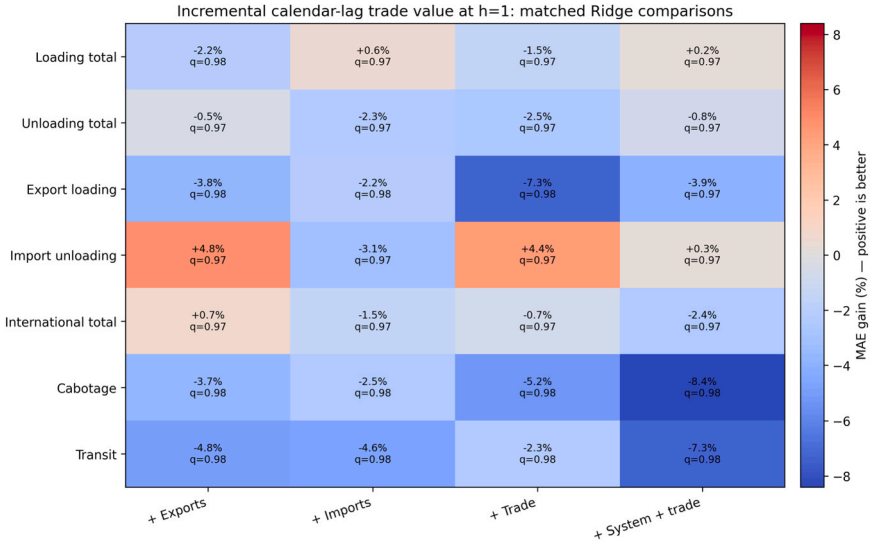


Fig. 3. Point Forecast Value Gain from calendar-lagged trade information at h= 1 using the same matched Ridge comparisons reported in Table 6. Cells show MAE gain and the BH-adjusted Clark–West q-value; no cell is FDR-supported at q < 0.05.

sensitivity checks.

The component forecasts in Table 5 use Ridge with alpha= 10 fixed at every expanding-window origin. The same penalty is retained in Table 6 and Supplementary Table S5, so adding predictors does not also change component regularization. Neither the component penalty nor the component model family is selected from the 28-month evaluation outcomes.

4.2.1. Target-representation robustness and model-family ranking

Changing the target to a seasonal log difference alters the model ranking, but the level-target reference remains the most accurate specification. Under Eq. (8), seasonal naive has test MASE of 1.106 at h= 1 and 1.108 at h= 3; benchmark statements therefore use test-window MAE. At h= 1, the best transformed-target result is Elastic Net at 66,607 TEU, 10.1% below the seasonal-naive MAE of 74,111 TEU. Extra Trees, Ridge, Random Forest, and SVR improve by only 0.5%–3.9%, while histogram gradient boosting is worse. At h= 3, Elastic Net is the only transformed family to improve on seasonal naive: 67,759 TEU, or 8.6% lower. The other transformed families are 7.6%–27.4% worse. Even the best transformed forecasts remain behind the level-target Ridge, whose MAE is 50,977 TEU at h= 1 and 54,110 TEU at h= 3. Supplementary Table S1 reports all transformed-target MAE and MASE values.

No single conclusion about algorithmic complexity follows from the transformed-target exercise. Several families improve on seasonal naive at h= 1, and Elastic Net does so at h= 3, but none reaches the accuracy of the level-target Ridge. The movement in rankings across target representations shows that this short national series is sensitive to how the

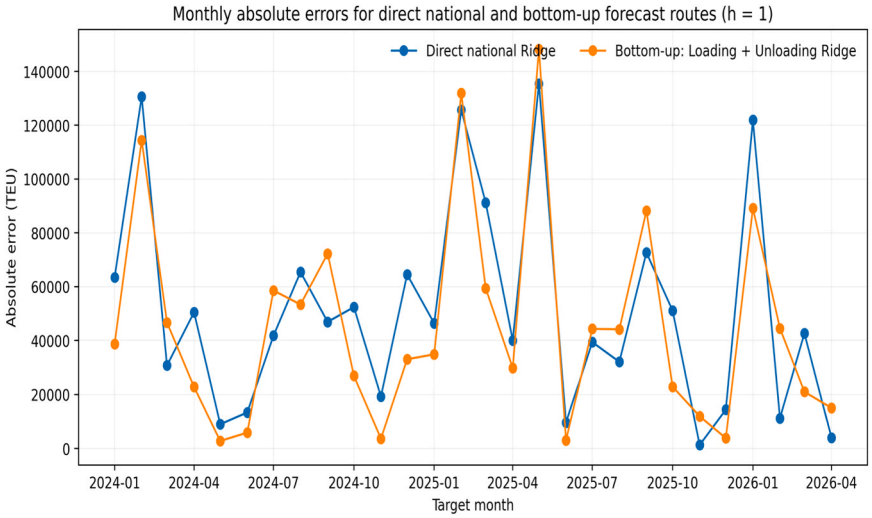


Fig. 4. Monthly absolute errors for the direct national Ridge forecast and the bottom-up aggregate at h= 1. The bottom-up route has lower mean MAE over the 28-month window, but the paired difference is not statistically supported (Wilcoxon p = 0.178).

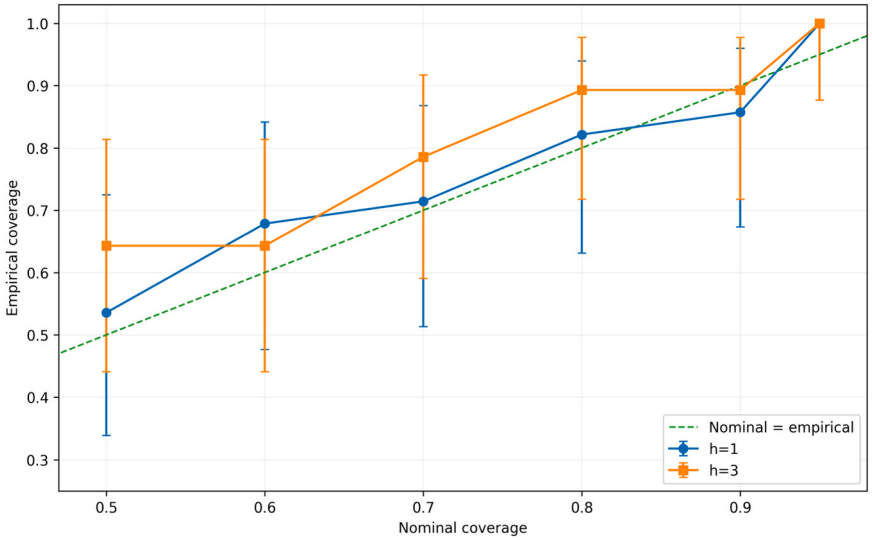


Fig. 5. Sequential same-horizon conformal coverage at h= 1 and h= 3 with binomial-reference 95% Clopper–Pearson intervals. Same-horizon residuals enter calibration only after their targets are observable.

prediction problem is posed.

4.3. Incremental value of calendar-lagged trade information

Each component-only Ridge in Table 6 is compared with four specifications that add calendar-lagged trade/system information. Clark-West is the primary test because the augmented models nest their baselines; Benjamini-Hochberg adjustment is then applied across the 28 Clark-West p-values. None of the 28 adjusted comparisons is FDR-supported at $q < 0.05$. The largest positive point gains occur for import unloading, but multiplicity adjustment leaves them descriptive. Supplementary Table S4 contains the complete robustness diagnostics for all 28 comparisons.

Negative point gains appear for cabotage in several specifications. Neither the Clark-West analysis nor the dependence-aware robustness checks, however, supports a confirmatory deterioration claim. Across the component results as a whole, calendar-lagged trade provides no reliable general improvement in forecast accuracy.

Positive gain indicates lower MAE than the matched component-only Ridge baseline. Confirmatory interpretation is based only on BH-adjusted Clark-West q-values.

4.4. Retrospective same-period trade information experiment

The same-period experiment looks at a different question. It examines what trade variables can add when the values for the same period are already known. It does not assume that these values were available at the real forecast time. According to ALFRED, they become available only after the forecast origin. For every component, the comparison starts from the fixed component-history Ridge reported in Table 5. Contemporaneous log exports and imports are then added, while the same expanding-window procedure is kept. Clark-West is used as the main test because the compared models are nested. The resulting tests across the seven components are adjusted using the Benjamini-Hochberg procedure. Loading total is the only FDR-supported case: MAE falls from 23,643 to 22,214 TEU, a 6.0% gain (Clark-West $p = 0.004$; BH $q = 0.027$). Export loading improves by 5.6% at the point-estimate level but is not FDR-supported ($p = 0.030$; $q = 0.105$). Supplementary Table S5 reports the complete experiment.

For loading total, the retrospective result establishes information

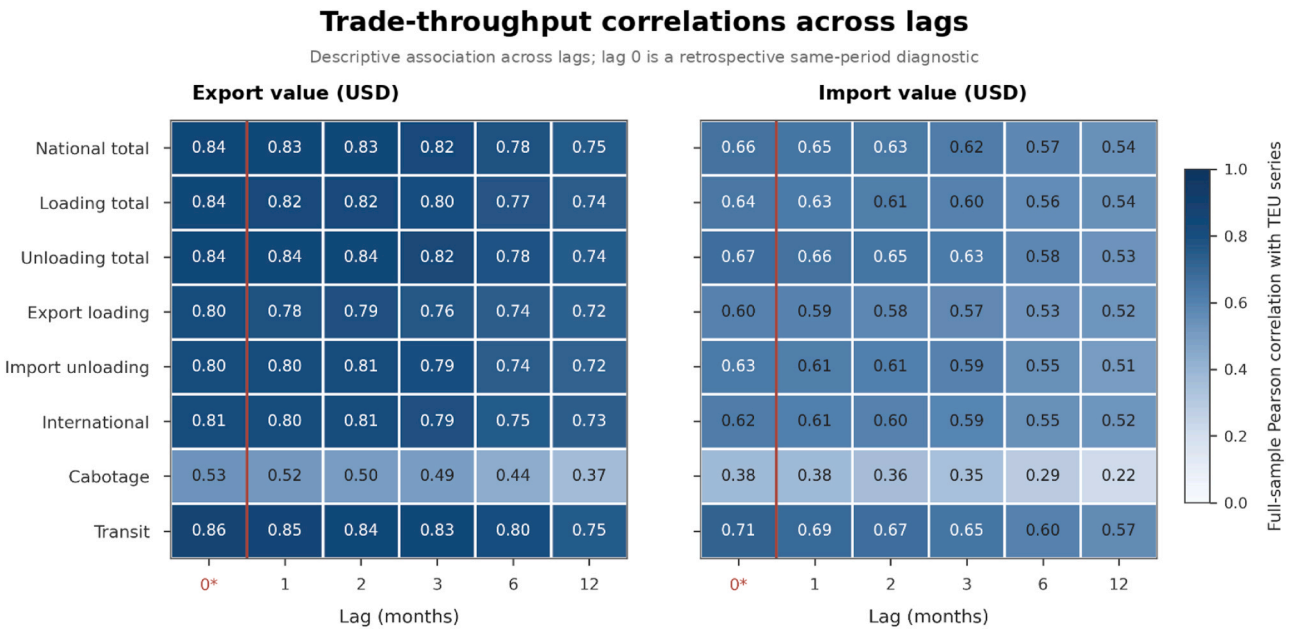


Fig. 6. Correlations between trade values and throughput series across lags 0–12 months. Lag 0 is shown only as a retrospective same-period diagnostic; correlation does not establish incremental forecast value.

Table 5
Fixed component-only Ridge forecasts by operational flow and horizon.

Component	h	Fixed family	Information set	MAE (TEU)	S-naive MAE	MAE gain	Unified MASE
Cabotage	1	Ridge	Component-only	7968	11,272	29.3%	0.850
Cabotage	3	Ridge	Component-only	8180	11,272	27.4%	0.876
Export loading	1	Ridge	Component-only	17,098	26,541	35.6%	0.701
Export loading	3	Ridge	Component-only	17,220	26,541	35.1%	0.707
Import unloading	1	Ridge	Component-only	18,916	30,213	37.4%	0.772
Import unloading	3	Ridge	Component-only	19,670	30,213	34.9%	0.803
International total	1	Ridge	Component-only	33,116	49,899	33.6%	0.732
International total	3	Ridge	Component-only	33,564	49,899	32.7%	0.742
Loading total	1	Ridge	Component-only	23,643	39,257	39.8%	0.678
Loading total	3	Ridge	Component-only	25,044	39,257	36.2%	0.720
Transit	1	Ridge	Component-only	17,886	36,543	51.1%	0.605
Transit	3	Ridge	Component-only	21,889	36,543	40.1%	0.755
Unloading total	1	Ridge	Component-only	26,713	40,437	33.9%	0.748
Unloading total	3	Ridge	Component-only	28,832	40,437	28.7%	0.811

content for the same reference month. It does not establish an operational gain because the trade release comes after the decision origin.

Seen together, the two trade experiments separate timing from information content. Calendar-lagged additions do not reduce MAE reliably at the operational origin. The same-period analysis detects current-month loading information only after that origin has passed.

4.5. Direct national and bottom-up aggregation-consistency check

At h= 1, the direct national route is the TEU-only Ridge in Table 4 and has MAE = 50,977 TEU. Summing the component-only Ridge forecasts for Loading total and Unloading total in Table 5 gives the bottom-up route, whose MAE is 45,411 TEU—10.9% below the direct forecast. The paired Wilcoxon test gives $p = 0.178$. The equal-weight combination has MAE = 48,051 TEU, 5.7% below direct national, with $p = 0.144$.

On average, the direct and bottom-up routes differ by approximately 1.55% of monthly throughput. This is used as an internal consistency check. The three MAE values are also reported descriptively for the direct, bottom-up, and equal-weight combined forecasts in Table 7.

4.6. Sequential same-horizon uncertainty intervals

The conformal calculation uses the same TEU-only Ridge point specification as Table 4. Pre-2024 rolling-origin residuals create separate warm-up pools by horizon. At the first 2024 test origin, 82 h= 1 residuals and 74 h= 3 residuals have matured; the pools later grow to 109 and 101. The warm-up history includes the high-volatility 2020–2021 observations.

The sequential same-horizon conformal interval results are summarized in Table 8. For h= 1, the 80% nominal interval covers 23 of 28 test months (82.1%) with an average half-width of 78,434 TEU; the corresponding binomial-reference 95% Clopper–Pearson interval for the observed coverage proportion is 63.1%–93.9%. The 95% interval covers all 28 months, has an average half-width of 140,153 TEU, and has a binomial-reference interval of 87.7%–100.0%. At h= 3, the 80% interval covers 25 of 28 months (89.3%); the average half-width is 115,019 TEU and the binomial-reference interval is 71.8%–97.7%. The 95% interval again covers all 28 months, with an average half-width of 173,497 TEU. These intervals summarize finite-sample coverage under a binomial reference and are not interpreted as exact time-series confidence intervals because the rolling coverage indicators may be serially dependent.

Table 6
Incremental calendar-lag trade value at h= 1: matched Ridge comparisons with Clark-West primary inference and BH-FDR control.

Component	Information set	Gain (%)	Clark-West p	BH q (28)	Interpretation
Loading total	Component + exports	-2.2	0.630	0.980	Not FDR-supported
Loading total	Component + imports	0.6	0.353	0.971	Not FDR-supported
Loading total	Component + trade	-1.5	0.508	0.971	Not FDR-supported
Loading total	Component + system + trade	0.2	0.328	0.971	Not FDR-supported
Unloading total	Component + exports	-0.5	0.437	0.971	Not FDR-supported
Unloading total	Component + imports	-2.3	0.448	0.971	Not FDR-supported
Unloading total	Component + trade	-2.5	0.358	0.971	Not FDR-supported
Unloading total	Component + system + trade	-0.8	0.149	0.971	Not FDR-supported
Export loading	Component + exports	-3.8	0.697	0.980	Not FDR-supported
Export loading	Component + imports	-2.2	0.719	0.980	Not FDR-supported
Export loading	Component + trade	-7.3	0.805	0.980	Not FDR-supported
Export loading	Component + system + trade	-3.9	0.501	0.971	Not FDR-supported
Import unloading	Component + exports	4.8	0.084	0.971	Not FDR-supported
Import unloading	Component + imports	-3.1	0.555	0.971	Not FDR-supported
Import unloading	Component + trade	4.4	0.122	0.971	Not FDR-supported
Import unloading	Component + system + trade	0.3	0.187	0.971	Not FDR-supported
International total	Component + exports	0.7	0.308	0.971	Not FDR-supported
International total	Component + imports	-1.5	0.548	0.971	Not FDR-supported
International total	Component + trade	-0.7	0.416	0.971	Not FDR-supported
International total	Component + system + trade	-2.4	0.408	0.971	Not FDR-supported
Cabotage	Component + exports	-3.7	0.956	0.980	Not FDR-supported
Cabotage	Component + imports	-2.5	0.878	0.980	Not FDR-supported
Cabotage	Component + trade	-5.2	0.920	0.980	Not FDR-supported
Cabotage	Component + system + trade	-8.4	0.921	0.980	Not FDR-supported
Transit	Component + exports	-4.8	0.980	0.980	Not FDR-supported
Transit	Component + imports	-4.6	0.794	0.980	Not FDR-supported
Transit	Component + trade	-2.3	0.794	0.980	Not FDR-supported
Transit	Component + system + trade	-7.3	0.927	0.980	Not FDR-supported

Table 7
Direct national, bottom-up aggregate, and simple combined forecasts at h= 1.

Approach	MAE (TEU)	Difference vs direct national	Wilcoxon p-value
Direct national forecast — TEU-only Ridge	50,977	Reference	—
Bottom-up aggregate — Loading + Unloading Ridge	45,411	10.9% lower	0.178
Equal-weight combination	48,051	5.7% lower	0.144

The reported half-width is a throughput-uncertainty margin measured in TEU. Its operational interpretation is bounded by the limitations discussed below.

5. Robustness and limitations

5.1. Lag-structure diagnostic

Trade and throughput are positively correlated at several full-sample lags, and the pattern varies only modestly across much of the lag range. A shared trend can create this appearance even when additional predictive value is small. For that reason, the FDR-controlled component tests are more informative about incremental forecasting value than the correlation map by itself.

Table 8
Sequential same-horizon conformal interval performance using the Table 4 TEU-only Ridge point model.

h	Nominal coverage	Empirical coverage	Binomial-reference 95% CP interval	Avg. half-width (TEU)
1	80%	82.1% (23/28)	[63.1%, 93.9%]	78,434
1	95%	100.0% (28/28)	[87.7%, 100.0%]	140,153
3	80%	89.3% (25/28)	[71.8%, 97.7%]	115,019
3	95%	100.0% (28/28)	[87.7%, 100.0%]	173,497

5.2. Target, model-selection, and inference robustness

With growth-compatible predictors and a seasonal-log-difference target, the best transformed model improves on the test-window seasonal-naive MAE by 10.1% at h= 1 and 8.6% at h= 3. Every transformed-target model remains less accurate than the level-target Ridge reference. Because origin-specific scaling puts seasonal-naive test MASE above 1, benchmark comparisons are made with MAE. The model ranking changes when the target representation is changed. This shows that results are sensitive to how the target is defined, although it still does not answer the broader question about model complexity. Supplementary Table S4 reports additional diagnostics for the component trade comparisons. These include dependence-aware maxT, paired-loss tests, Wilcoxon tests and bootstrap results. The findings are consistent with the main interpretation obtained from the BH-adjusted Clark-West tests.

5.3. Data-vintage, holdout, and uncertainty limitations

The limitations mainly come from three parts: what information was actually observable at the forecast date, how the evaluation sample was used during the analysis, and what national TEU uncertainty can indicate. For the UAB releases examined in 2026, data were published around 10–15 days after month-end [1]. The same release timing was not reconstructed for earlier years, so it may have been different. In this study, the forecast origin is the UAB publication date for the latest available throughput month. Under this definition, the h= 1 prediction is produced after the target calendar month has begun. It is therefore a current-month nowcast in operational terms, not a forecast available before the start of the target month. The h= 3 horizon remains forward-looking beyond the current month. For the trade variables, ALFRED gives information about revisions and data vintages. A complete point-in-time panel, however, was not rebuilt. The trade-augmented results should therefore be read as pseudo-real-time analyses using calendar lags, rather than a fully reconstructed real-time forecasting exercise. The archived export view dated 15 May 2026 stops at March 2026. In the update dated 15 June, April 2026

appears. A similar timing is shown for imports, where the public page also connects April with the 15 June update. Still, the exact first-vintage timestamp for imports was not independently reconstructed.

The length and use of the comparison sample is another limitation. The evaluation window contains only 28 months and it was used repeatedly during model development and sensitivity checks. It was therefore not kept as a completely untouched holdout sample. The small number of observations also matters for the conformal results. With $n = 28$, the binomial-reference coverage intervals are wide. In addition, serial dependence among rolling coverage indicators means that these intervals should be viewed as descriptive reference intervals rather than exact time-series confidence intervals. Any distinction made between calmer and more variable periods should for this reason be considered descriptive rather than a separate validation result. Unified MASE also needs some caution. Its denominator is based on an expanding training history, so highly volatile periods inside the training sample, particularly 2020–2021, can affect the scale used at later origins. This is why MAE calculated directly over the test window and comparisons against the seasonal-naive benchmark are reported beside MASE rather than relying on MASE alone.

Another issue is that the trade series and throughput series do not have the same seasonal treatment. Exports and imports are seasonally adjusted, but the throughput target is not. Final-vintage seasonal adjustment can itself change when later observations become available. This creates some additional possibility of look-ahead. It also removes seasonal movements from the predictors while these movements are still present in the target. This mismatch can be one reason why the additional contribution of trade variables remains limited. The national TEU series has another practical restriction. Monthly national totals do not include terminal operating conditions, and for this reason uncertainty around national throughput cannot directly be translated into physical capacity needs at terminal level.

A stronger external validation would require a complete point-in-time reconstruction of the ALFRED vintages and, preferably, a later evaluation period that was never used during model development. A longer national time series, data from several ports, and trade variables expressed in physical volume could also extend the evidence. For using forecast uncertainty directly in terminal planning, local operational information would still be needed. This includes dwell time, occupancy, berth and crane productivity, and staffing data.

6. Operational implications

The operational meaning of the two horizons differs. The $h = 1$ result should be interpreted as current-month nowcasting performance because the forecast origin occurs after the target month has begun. It therefore does not represent a forecast available before the start of that month. The $h = 3$ result retains a forward forecasting interpretation because its target lies beyond the current calendar month at the decision origin. For routine monthly forecasting, the operational-history model is the practical reference in this dataset. Under the observed release schedules, a month's trade values arrive after both the forecast origin and the UAB throughput release for that month, so their contemporaneous content is available too late to alter the decision at that origin.

Mean error alone can overstate how typical an improvement is. Here, several large seasonal-naive errors contribute materially to the mean Ridge advantage; the median absolute-error comparison is therefore useful alongside the mean.

Target representation deserves the same caution. The level-target Ridge is more accurate than the transformed-target alternatives in this sample, but the changes in model order across representations do not settle the broader question of model complexity.

The trade evidence is best read as a release-timing result. Calendar-lagged values do not yield a reliable forecasting gain, yet the retrospective same-period experiment finds useful information for loading. That information would become operationally relevant only if trade were released before the corresponding throughput statistics.

Throughput-uncertainty margins can support demand-contingency planning, but a TEU half-width is not berth, yard, equipment, or staffing capacity. Ports would need local dwell-time, occupancy, productivity, peak-demand, and service-level parameters before converting those margins into resource requirements.

The direct national model and the bottom-up approach give two different ways of looking at monthly throughput. The national Ridge forecast is kept as the main reference. The bottom-up total is used mainly to check whether the component-level results are reasonably consistent with the national series.

From a practical side, the results support using a release-aware history model as the main forecasting base. Before moving to more complex models, it is also useful to check whether the result changes when the target is represented in a different way. Point forecasts should not be reported alone, and throughput uncertainty margins should be shown beside them. This reading is also in line with recent transport-management studies that emphasize temporal validity in forecasting and the need to include uncertainty when forecasts are used for decisions [20,21].

7. Conclusion

For Türkiye's national container series, the level-target operational-history Ridge in Table 4 produces the lowest forecast error among the national specifications considered. The average improvement over seasonal naive, however, does not occur similarly across all months. A small number of months with large seasonal-naive errors account for an important part of the mean difference. For this reason, the size of the advantage is different when mean and median errors are compared. The evaluation period is also short, so the estimated average gain still has considerable uncertainty around it.

Calendar-lagged trade does not reduce national MAE, and none of the 28 component-level additions is FDR-supported. At $h = 1$, the national Clark–West test rejects despite higher MAE, but the result is not replicated at $h = 3$. The retrospective same-period experiment contains one FDR-supported result, for loading total. This shows current-month trade information in the data, but the information enters after the defined decision origin. Changing target representation reorders the model families without displacing the level-target Ridge as the most accurate evaluated specification. Sequential same-horizon conformal calibration complements the point forecasts with empirical coverage estimates and binomial-reference coverage intervals for the remaining throughput uncertainty.

Operationally, the $h = 1$ results describe current-month nowcasting rather than forecasting before the target month begins, while $h = 3$ retains a forward multi-month interpretation. For port planning, the evidence favors release-aware operational-history forecasting and throughput-uncertainty margins for demand-contingency planning. The same-period trade signal would become operationally usable only with an earlier release schedule. Later external data, reconstructed real-time vintages, and port-specific operating information are needed to assess how far these results generalize beyond the national series.

CRediT authorship contribution statement

Mehdi Safaei: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Ethics statement

The analysis uses publicly available aggregate monthly transport and trade statistics and includes no human participants, animal subjects, experiments, surveys, interviews, or personally identifiable information.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation and revision of this work, the author used AI (Ginger and Grammarly) for language checking, grammar review, readability, and formatting consistency. The author reviewed and edited all resulting suggestions as needed, verified the scientific content, data, analyses, citations, and references, and takes full responsibility for the content of the publication.

Funding

The author received no external funding for this research.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The author has no acknowledgements to declare.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.team.2026.10.005.

Data availability

Data will be made available on request.

References

[1] UAB, 2026, Aylık Yayınlar [Monthly publications]. https://denizcilik.uab.gov.tr/aylik-yayinlar.
[2] B. Lamdjad, Comparative forecasting of container throughput for maritime logistics using statistical, machine learning and explainable AI models, Marit. Bus. Rev. 11 (3) (2026) 202–218, https://doi.org/10.1108/MABR-08-2025-0086.
[3] M. Milenković, N. Milosavljević, N. Bojović, S. Val, Container flow forecasting through neural networks based on metaheuristics, Oper. Res. 21 (2) (2021) 965–997, https://doi.org/10.1007/s12351-019-00477-1.
[4] S. Shankar, S. Punia, P.V. Ilavarasan, Deep learning-based container throughput forecasting: A triple bottom line approach, Ind. Manag. Data Syst. 121 (10) (2021) 2100–2117, https://doi.org/10.1108/IMDS-12-2020-0704.
[5] N.D. Tan, H.C. Yu, L.N.B. Long, S.-S. You, Time series forecasting for port throughput using recurrent neural network algorithm, J. Int. Marit. Saf. Environ. Aff. Shipp. 5 (4) (2021) 175–183, https://doi.org/10.1080/25725084.2021.2014245.
[6] J. Wang, Y. Shao, H. Jiang, Y. An, A multi-variable hybrid system for port container throughput deterministic and uncertain forecasting (Article), Expert. Syst. Appl. 237 (2024) 121546, https://doi.org/10.1016/j.eswa.2023.121546.
[7] G. Wu, Y.-C. Hu, Y. Chiu, P. Jiang, R. Chi, Forecast combination using grey relational analysis and Choquet fuzzy integral for container throughput forecasting (Article), Expert. Syst. Appl. 252 (2024) 124170, https://doi.org/10.1016/j.eswa.2024.124170.
[8] S. Xu, S. Zou, J. Huang, W. Yang, F. Zeng, Comparison of different approaches of machine learning methods with conventional approaches on container throughput forecasting (Article), Appl. Sci. 12 (19) (2022) 9730, https://doi.org/10.3390/app12199730.
[9] M. Eskafi, M. Kowsari, A. Dastgheib, G.F. Ulfarsson, P. Taneja, R.I. Thorarinsdottir, Mutual information analysis of the factors influencing port throughput, Marit. Bus. Rev. 6 (2) (2021) 129–146, https://doi.org/10.1108/MABR-05-2020-0030.
[10] D. Morales-Ramírez, M.D. Gracia, J. Mar-Ortiz, Forecasting national port cargo throughput movement using autoregressive models (Article), Case Stud. Transp. Policy 19 (2025) 101322, https://doi.org/10.1016/j.cstp.2024.101322.
[11] F. Zeng, S. Xu, A hybrid container throughput forecasting approach using bi-directional hinterland data of port (Article), Sci. Rep. 14 (2024) 25502, https://doi.org/10.1038/s41598-024-77376-9.
[12] H. Hewamalage, K. Ackermann, C. Bergmeir, Forecast evaluation for data scientists: Common pitfalls and best practices, Data Min. Knowl. Discov. 37 (2) (2023) 788–832, https://doi.org/10.1007/s10618-022-00894-5.
[13] C. Bergmeir, J.M. Benítez, On the use of cross-validation for time series predictor evaluation, Inf. Sci. 191 (2012) 192–213, https://doi.org/10.1016/j.ins.2011.12.028.
[14] L.J. Tashman, Out-of-sample tests of forecasting accuracy: an analysis and review, Int. J. Forecast. 16 (4) (2000) 437–450, https://doi.org/10.1016/S0169-2070(00)00065-0.
[15] D. Hopp, Benchmarking econometric and machine learning methodologies in nowcasting, in: UNCTAD Research Paper No, 83, United Nations Conference on Trade and Development, 2022, https://doi.org/10.18356/27082814-83.
[16] M. Chinn, B. Meunier, S. Stumpner, Nowcasting world trade with machine learning: A three-step approach, J. Forecast. 45 (6) (2026) 2954–2968, https://doi.org/10.1002/for.70165.
[17] Z.H. Munim, C.S. Fiskin, B. Nepal, M.M.H. Chowdhury, Forecasting container throughput of major Asian ports using the Prophet and hybrid time series models, Asian J. Shipp. Logist. 39 (2) (2023) 67–77, https://doi.org/10.1016/j.ajsl.2023.02.004.
[18] G. Shafer, V. Vovk, A tutorial on conformal prediction, J. Mach. Learn. Res. 9 (2008) 371–421.
[19] V. Vovk, A. Gammerman, G. Shafer, Algorithmic learning in a random world, Springer, 2005.
[20] R.C. Dela Cruz, N.I. Sore, Forecast accuracy versus operational trigger value in next-day truck utilization forecasting: evidence from Philippine logistics, Transp. Econ. Manag. 4 (2026) 152–173, https://doi.org/10.1016/j.team.2026.07.001.
[21] D. De Santis, M. Giagnorio, G. Matteucci, Investments in local public transport: a systematic literature review on evaluation methods with a focus on the role of uncertainty, Transp. Econ. Manag. 3 (2025) 441–459, https://doi.org/10.1016/j.team.2025.11.002.
[22] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (4) (2006) 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.
[23] F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, J. Bus. Econ. Stat. 13 (3) (1995) 253–263, https://doi.org/10.1080/07350015.1995.10524599.
[24] D. Harvey, S. Leybourne, P. Newbold, Testing the equality of prediction mean squared errors, Int. J. Forecast. 13 (2) (1997) 281–291, https://doi.org/10.1016/S0169-2070(96)00719-4.
[25] T.E. Clark, K.D. West, Approximately normal tests for equal predictive accuracy in nested models, J. Econ. 138 (1) (2007) 291–311, https://doi.org/10.1016/j.jeconom.2006.05.023.
[26] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B 57 (1) (1995) 289–300, https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
[27] R.J. Hyndman, R.A. Ahmed, G. Athanasopoulos, H.L. Shang, Optimal combination forecasts for hierarchical time series, Comput. Stat. & Data Anal. 55 (9) (2011) 2579–2589, https://doi.org/10.1016/j.csda.2011.03.006.
[28] S.L. Wickramasuriya, G. Athanasopoulos, R.J. Hyndman, Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization, J. Am. Stat. Assoc. 114 (526) (2019) 804–819, https://doi.org/10.1080/01621459.2018.1448825.
[29] K. Sanguri, S. Shankar, S. Punia, S. Patra, Hierarchical container throughput forecasting: The value of coherent forecasts in the management of ports operations (Article), Comput. & Ind. Eng. 173 (2022) 108651, https://doi.org/10.1016/j.cie.2022.108651.
[30] Federal Reserve Bank of St. Louis, International Merchandise Trade Statistics: Exports: Commodities for Turkiye [XTEXVA01TRM667S, ALFRED, Federal Reserve Bank of St. Louis, 2026.
[31] Federal Reserve Bank of St. Louis, ALFRED. International Merchandise Trade Statistics: Imports: Commodities for Turkiye [XTIMVA01TRM667S, Federal Reserve Bank of St. Louis, 2026.
[32] J.H. Friedman, Greedy function approximation: A gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232, https://doi.org/10.1214/aos/1013203451.
[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[34] W.K. Newey, K.D. West, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, Econometrica 55 (3) (1987) 703–708, https://doi.org/10.2307/1913610.
[35] C.J. Clopper, E.S. Pearson, The use of confidence or fiducial limits illustrated in the case of the binomial, Biometrika 26 (4) (1934) 404–413, https://doi.org/10.1093/biomet/26.4.404.
[36] UAB, 2023, February 13, Ulaştırma ve Altyapı Bakanlığı: Deprem bölgesinde tüm ekiplerimizin çalışmaları sürüyor. https://sgb.uab.gov.tr/haberler/ulastirma-ve-altyapi-bakanligi-deprem-bolgesinde-tum-ekiplerimizin-calismalari-suruyor.