logistics

MDPI

Article

Forecasting Port Access Traffic Under Temporal Heterogeneity: A Leakage-Free Evaluation of Conditioning and Fixed Model Assignment

Bechir Ben-Daya [1,*] and Jean-François Audy [1,2]

[1] Department of Management, Université du Québec à Trois-Rivières, Trois-Rivières, QC G8Z 4M3, Canada; jean-francois.audy@uqtr.ca
[2] Interuniversity Research Center on Enterprise Network, Logistics and Transportation (CIRRELT), Québec, QC G1V 0A6, Canada
* Correspondence: bechir.ben.daya@uqtr.ca

Abstract

Background: Port access traffic combines pronounced calendar structure with substantial day-to-day variability, raising the question of whether temporal heterogeneity should condition a common forecasting configuration or support model specialization. *Methods*: Daily car and freight-truck arrivals at an urban non-containerized port were forecast at 1-, 7-, and 14-day horizons using statistical, machine-learning, and composite models under a leakage-free rolling-origin protocol. A unified calendar-conditioned ensemble was compared with a prespecified regime-based model assignment, with development, test, and temporal stress-test periods kept distinct. *Results*: The unified ensemble achieved next-day R^2 values of 0.75 for trucks and 0.80 for cars on the held-out 2018 test period and 0.73 and 0.75, respectively, under the 2019 temporal stress-test. Model preference varied jointly with temporal regime, forecast horizon, and evaluation period. Although fixed assignment improved car accuracy on the 2018 test period, the model–regime dominance required to justify it was not sufficiently stable across horizons and evaluation periods. *Conclusions*: Calendar conditioning provides a robust strategy for the present forecasting problem. More generally, fixed regime-based specialization should be supported by sufficiently stable model–regime dominance rather than inferred from identifiable temporal heterogeneity alone.

Keywords: port access traffic; daily demand forecasting; temporal heterogeneity; regime-based forecasting; rolling-origin evaluation; SARIMAX; recurrent neural networks; ensemble forecasting

Check for updates

Academic Editor: Robert Handfield

Received: 2 July 2026
Revised: 2 September 2026
Accepted: 3 September 2026
Published: 8 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Port access gates form a critical interface between inland traffic and port operations, where variations in vehicle arrivals affect gate workload, resource allocation, and short-term capacity requirements [1]. At urban, non-containerized ports, these flows also interact with the surrounding road network, making their temporal distribution relevant to both port operations and urban congestion [2]. Reliable anticipation of daily access demand can therefore support short-term operational planning and coordination with local stakeholders [3]. This study examines the Port of Trois-Rivières, Quebec, Canada, using a continuous five-year record of gate-level vehicle access from 2015 to 2019.

Daily port access traffic exhibits pronounced temporal structure, dominated in this case by a weekday–weekend contrast: weekday volumes are generally higher and more variable, whereas weekend volumes are lower and more regular (Section 3). Because the calendar regime of a future target date is known at the forecast origin, this heterogeneity provides information that can be exploited without look-ahead. It also raises a modeling question: should regime information condition a common forecasting configuration, or does it support assigning different models to different regimes? This question is examined at three forecast horizons: one day, one week, and two weeks (h = 1, 7, 14), corresponding to next-day gate operations, weekly scheduling, and two-week planning. Next-day accuracy is the primary operational target, while the longer horizons characterize how forecasting skill changes with lead time.

Addressing this forecasting problem raises three interconnected dimensions: a scientific question concerning how observable temporal heterogeneity should inform model specialization, a methodological requirement for temporally valid and information-consistent model comparison, and an applied forecasting problem concerning daily port access demand at operational horizons. Comparisons across statistical, machine-learning, and composite models require an evaluation design that preserves temporal order and restricts predictors to information available at each forecast origin. At the same time, identifiable calendar heterogeneity raises the question of how regime information should be incorporated into the forecasting strategy, including whether it should condition a common configuration or support model specialization across regimes. These questions are particularly relevant to daily port access traffic, where the forecasting target, operational horizons, and available information differ from those of more aggregate port-demand applications. Section 2 examines these issues in the forecasting literature and establishes the specific positioning of the present study.

Accordingly, this study has three objectives: (O1) to compare statistical, machine-learning, and composite forecasting models for daily port access traffic under a common leakage-free rolling-origin protocol; (O2) to characterize how model performance varies across the weekday–weekend regime and forecast horizons; and (O3) to identify and evaluate a retained forecasting configuration out of sample, with next-day forecasting as the primary operational target. These objectives lead to three research questions: (Q1) how does the relative accuracy of the compared model families vary across temporal regimes and forecast horizons? (Q2) Does assigning models by regime improve accuracy over a unified conditioned forecasting configuration out of sample, and is any such improvement sufficiently stable across forecast horizons and evaluation periods? (Q3) How robust is the retained configuration under anomalous traffic conditions, and how does its forecasting skill degrade as the horizon lengthens?

The study makes three contributions. First, scientifically, it shows that identifiable temporal heterogeneity is not equivalent to stable model–regime dominance and formulates a falsifiable condition for fixed regime-based forecasting: model assignment requires sufficiently stable model–regime dominance across forecast horizons and evaluation periods. Second, methodologically, it provides a leakage-free rolling-origin comparative framework that aligns forecasts with their target dates, restricts predictors to information available at the forecast origin, separates development, test, and stress-test periods, and combines regime-specific and multi-horizon reporting with formal significance testing. This framework makes the proposed stability condition empirically testable without train–test contamination. Third, in applied terms, the study identifies a unified calendar-conditioned ensemble that maintains strong next-day held-out performance and degrades gracefully under the 2019 temporal stress-test, while explicitly documenting horizon-dependent skill decay and statistically detectable weekly dependence remaining in the car residuals.

The empirical analysis is based on the Port of Trois-Rivières, for which a continuous five-year record of gate-level vehicle entry and exit timestamps by vehicle type is available. This granularity enables daily access demand to be reconstructed directly from individual gate records and supports temporally ordered out-of-sample evaluation. The empirical findings and reported accuracy levels remain specific to this facility and observation period. Their transferability to other ports or access-controlled systems is therefore treated as a hypothesis requiring external validation rather than as a demonstrated result.

The remainder of the paper is organized as follows. Section 2 reviews port and logistics traffic forecasting, statistical, learning-based, and hybrid approaches, regime-dependent modeling, forecasting evaluation, and recent deep-learning architectures. Section 3 describes the data, admissible information set, forecasting models, and leakage-free rolling-origin protocol. Section 4 reports predictive performance, regime-specific behavior, residual diagnostics, and the 2019 temporal stress-test. Section 5 discusses the implications of the findings, robustness, operational relevance, transferability, and limitations. Section 6 concludes.

2. Literature Review

2.1. Forecasting Port Access and Logistics Traffic

Vehicle flows at port access points directly affect gate operations and their interaction with surrounding transport networks. Previous studies have examined this interface primarily through gate-congestion and queueing models, showing how variations in truck arrivals and service conditions influence waiting times and terminal performance [4,5]. At a broader planning level, traffic forecasting supports port capacity and resource decisions [6], while machine-learning and optimization approaches are increasingly used to improve seaport operational efficiency [7]. These studies establish demand anticipation as an important component of port planning and operations, but forecasting vehicle access itself remains a more specialized problem.

A smaller body of research addresses vehicle or traffic flows at resolutions closer to port access operations. Jin et al. [8] forecast daily gate-in and gate-out volumes at a container terminal using machine-learning methods, providing a direct precedent for daily gate-level forecasting. Ali Nadi et al. [9] predict short-term outbound truck traffic at the Port of Rotterdam using information exchanged among logistics hubs, illustrating the value of operational information for port-related traffic prediction. At a different operational interface, Yu et al. [10] address fine-grained vessel traffic-flow prediction in port waters by integrating graph structure and time-series information. Beyond port access specifically, data-driven forecasting combining statistical and deep-learning approaches has been applied to transportation demand across distribution centers [11] and to material demand in manufacturing [12]. Together, these studies demonstrate the relevance of learning-based forecasting for short-term logistics demand while also showing the diversity of prediction targets, temporal resolutions, and information environments considered.

A related stream focuses on aggregate port throughput rather than gate-level vehicle access. Rashed et al. [13] use an ARIMA-intervention model for short-term container-throughput forecasting at the Port of Antwerp, while Farhan and Ong [14] apply seasonal ARIMA models to container throughput. Such studies demonstrate the continued relevance of explicit seasonal structure in port forecasting, but their aggregate throughput targets differ from the daily vehicle-arrival process considered here. Gate-level access forecasting requires predicting the temporal demand presented directly to the port–land interface, where pronounced calendar effects and short operational horizons can play a central role.

Overall, the literature supports both statistical and learning-based approaches to port and logistics forecasting, while revealing substantial diversity in forecasting targets,

temporal resolutions, and information environments. Daily gate-level vehicle forecasting remains a comparatively specialized setting, distinct from aggregate throughput forecasting and characterized by short operational horizons and pronounced calendar structure.

2.2. Statistical, Learning-Based, and Hybrid Models

Time-series forecasting encompasses model families that represent temporal structure and nonlinear relationships in different ways. Seasonal statistical models such as Seasonal Autoregressive Integrated Moving Average (SARIMA) and Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX) explicitly capture autoregressive dependence and recurring seasonal patterns and remain widely used in port and logistics forecasting [13–15]. Learning-based approaches include tree-based ensemble methods, support-vector methods, feed-forward neural networks, and recurrent architectures, among others [8,9,16–18]. Within these families, gradient boosting constructs predictive functions through ensembles of decision trees, while recurrent neural networks such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) learn sequential dependencies through internal states. Comparative evidence indicates that relative performance varies across datasets and forecast horizons [19], suggesting that model suitability depends on the characteristics of the forecasting problem rather than on model family alone.

Hybrid, decomposition-based, and forecast-combination approaches build on this complementarity in different ways. Decomposition-based methods separate a time series into components that can be modeled individually before recombination, while hybrid architectures combine models with different representational properties and forecast combinations aggregate predictions from multiple models. These strategies are represented in port and logistics forecasting by decomposition–ensemble approaches [8], sequential SARIMA–LSTM and SARIMAX–ANN architectures [16,20], broader hybrid frameworks [21], and linear or nonlinear forecast combinations [18]. Their common rationale is that different models or components may capture complementary structures within the same time series. Reported gains, however, vary across applications, reinforcing the view that complementarity is an empirical property to be established rather than an inherent advantage of model combination.

This context makes evaluation design central to comparative forecasting. Because time-series observations are temporally ordered, estimates of predictive performance depend on how the transition from observed history to future targets is reproduced during evaluation. Out-of-sample and rolling-origin approaches preserve this temporal structure and have therefore become important foundations of forecast evaluation [22–24]. Related research has also clarified the conditions under which cross-validation can be valid for autoregressive forecasting [25] and the risks created when information unavailable at prediction time enters model development or evaluation [26,27].

Beyond point estimates of error, comparative forecasting also requires distinguishing systematic differences in predictive accuracy from sample variation. The Diebold–Mariano framework provides a formal basis for comparing forecast errors [28], with subsequent corrections addressing finite-sample properties [29]. Taken together, this literature shows that comparisons among statistical, learning-based, and composite models depend not only on their predictive architectures but also on temporally valid evaluation and statistically grounded comparison.

2.3. Temporal Heterogeneity and Regime-Dependent Forecasting

Temporal heterogeneity has long motivated forecasting approaches in which predictive relationships vary across states or regimes. Markov-switching models represent such

variation through state-dependent parameters and probabilistic transitions between latent regimes [30]. Mixture-of-experts models provide another form of specialization by combining multiple predictive components through a gating mechanism that determines their contribution according to the observed input or inferred state [31]. Forecast-combination research addresses a related problem from a different perspective by aggregating predictions from multiple models to exploit complementary predictive information [32,33]. Although these approaches differ in formulation, they collectively show that temporal heterogeneity can be accommodated through state-dependent dynamics, specialized predictive components, or combinations of forecasts.

Temporal structure can also be represented through observable calendar information. Day-of-week, weekend, holiday, and other recurring calendar effects can be incorporated directly as predictors when they are known for the forecast target date [34]. Unlike latent-state approaches, this representation does not require the temporal state to be inferred from the observed series. Calendar information can therefore provide an ex ante representation of recurring temporal variation within the forecasting information set.

These different approaches illustrate that identifiable temporal heterogeneity does not prescribe a unique forecasting architecture. Temporal variation may be represented through model dynamics, specialized components, forecast combinations, or observable predictors, while comparative forecasting evidence indicates that relative model performance can vary across datasets and forecast horizons [19]. The relationship between an observed temporal structure and the forecasting architecture best suited to exploit it therefore remains an empirical modeling question.

For daily demand and access-traffic forecasting, this question becomes particularly relevant when recurring calendar patterns distinguish subsets of observations with markedly different statistical characteristics. Determining whether such observable temporal structure should condition a common predictive configuration or support model specialization therefore requires explicit comparative evaluation.

2.4. Recent Deep-Learning Architectures

Recent deep-learning architectures have substantially expanded the methodological landscape of time-series forecasting. Transformer-based models have introduced different mechanisms for representing long-range temporal dependencies and improving computational efficiency, including sparse attention in Informer [35], decomposition and auto-correlation in Autoformer [36], patch-based representations in PatchTST [37], and inverted tokenization in iTransformer [38]. Related architectures such as TimesNet represent temporal variation through a two-dimensional formulation [39], while Temporal Fusion Transformers combine recurrent processing, attention, and variable-selection mechanisms for multi-horizon forecasting [40]. Non-transformer architectures such as NHITS have likewise demonstrated strong performance through hierarchical interpolation and multi-rate signal sampling [41].

Increasing architectural sophistication, however, does not imply systematic superiority across forecasting settings. Zeng et al. [42] show that simple linear models can match or outperform several transformer-based approaches on widely used long-term forecasting benchmarks, illustrating the importance of aligning model complexity with the characteristics of the forecasting problem. More generally, comparative performance depends on the data structure, forecast horizon, information set, and evaluation design rather than on model class alone.

The present setting differs from the large-scale, long-sequence, and often multivariate forecasting problems for which many recent architectures have been developed and benchmarked. It consists of two univariate daily series, approximately three years of development

data, pronounced calendar structure, and three fixed operational horizons. A controlled comparison with substantially higher-capacity architectures would require architecture-specific tuning and regularization while maintaining comparable model-selection budgets and the same leakage-free information constraints used for the models evaluated here. Such a benchmark constitutes a distinct experimental extension. The present study therefore focuses on statistical, recurrent, tree-based, and composite configurations suited to its comparative question, while identifying evaluation of recent deep-learning architectures under the same rolling-origin protocol as an avenue for further research.

2.5. Synthesis and Positioning

The reviewed literature establishes three considerations relevant to the present forecasting problem. First, port and logistics forecasting encompasses diverse targets, temporal resolutions, and information environments. Statistical, learning-based, and composite approaches have all demonstrated predictive value, but daily gate-level vehicle access remains a comparatively specialized forecasting target, distinct from aggregate port throughput and other logistics-demand applications. Second, meaningful comparison across forecasting models depends on temporally valid out-of-sample evaluation, control of the information available at each forecast origin, and appropriate statistical comparison of competing forecasts. Third, temporal heterogeneity can be accommodated in several ways, including state-dependent dynamics, specialized predictive components, forecast combinations, and observable calendar predictors. The existence of identifiable temporal structure therefore does not, by itself, prescribe how that structure should be incorporated into a forecasting system.

Against this background, a specific unresolved question concerns observable temporal regimes known at the forecast origin. Such information can condition a common forecasting configuration or support specialization across regimes, but these alternatives imply different assumptions about model–regime relationships. Fixed assignment, in particular, presumes that the predictive advantage motivating a model–regime correspondence is sufficiently persistent to support that assignment beyond the setting in which it is identified. The port and logistics forecasting studies summarized in Table 1 do not directly evaluate this conditioning-versus-assignment question while jointly combining temporally held-out evaluation, multiple forecast horizons, and formal forecast comparison.

The present study addresses this intersection rather than proposing a new forecasting model family. It compares statistical, learning-based, and composite models for daily port access traffic under a common leakage-free rolling-origin protocol and examines how their relative performance varies across observable weekday–weekend regimes and forecast horizons. It then directly compares a unified calendar-conditioned configuration with a prespecified fixed regime-based assignment and examines whether the model–regime dominance required to support fixed assignment remains sufficiently stable across development and test periods. Table 1 positions the experimental design relative to representative port and logistics forecasting studies in terms of forecasting target, temporal resolution, regime treatment, temporally held-out evaluation, multi-horizon assessment, and formal forecast comparison.

Table 1 highlights the distinctive combination examined in the present study: daily gate-level forecasting, explicit calendar-regime treatment, temporally held-out rolling-origin evaluation, multi-horizon assessment, and formal forecast comparison within a common experimental design.

Table 1. Comparison of representative port-related forecasting studies and the present study.

Study	Forecasting Target/Setting	Temporal Resolution	Modeling Approach	Explicit Regime Treatment	Temporally Held-out Evaluation	Multi-Horizon Evaluation	Formal Forecast Comparison
Jin et al. [8]	Container gate-in/out, container terminal	Daily	XGBoost decomposition–ensemble; ARIMA benchmark	No	Yes	No	No
Ali Nadi et al. [9]	Outbound truck traffic, Port of Rotterdam	Hourly	FFNN with evolutionary feature selection using logistics-hub information	No	Yes	No	No
Rashed et al. [13]	Container throughput, Port of Antwerp	Monthly	SARIMA, SARIMA-intervention, ARIMAX	No	Yes	No	No
Yang & Chang (2020) [17]	Container throughput, five Taiwanese ports	Monthly	CNN–LSTM; LR, AdaBoost, RFR, and SVR benchmarks	No	Yes	No	Yes (Wilcoxon signed-rank and Friedman tests)
Fiskin et al. [16]	Domestic shipping market, Turkey	Monthly	SARIMAX, MLP, LSTM, NARX, and SARIMAX–ANN hybrid	No	Yes	No	Yes (Diebold–Mariano and paired *t*-tests)
Wu et al. (2024) [18]	Container throughput, eight Chinese ports	Quarterly	Statistical, ML, and DL models with linear and nonlinear forecast combinations	No	Yes	Yes (1–4 quarters ahead)	No
This study	Car and freight-truck gate arrivals, urban non-containerized port	Daily	Statistical, tree-based, recurrent, and composite models	Calendar conditioning vs. fixed model assignment	Yes—leakage-free rolling origin	Yes (h = 1, 7, 14)	Yes (Diebold–Mariano; Wilcoxon signed-rank where warranted)

Abbreviations used in this table: XGBoost, Extreme Gradient Boosting; FFNN, Feed-Forward Neural Network; CNN, Convolutional Neural Network; LR, Linear Regression; RFR, Random Forest Regression; SVR, Support Vector Regression; MLP, Multilayer Perceptron; NARX, Nonlinear Autoregressive model with Exogenous Inputs; ML, Machine Learning; DL, Deep Learning.

3. Methods

3.1. Forecasting Problem and Evaluation Protocol

This study addresses short- and medium-term forecasting of daily port access traffic for freight trucks and cars at the Port of Trois-Rivières. The prediction target is the daily count of vehicle arrivals of each type, indexed by entry timestamp. Forecasts are produced at three horizons, $h \in \{1, 7, 14\}$ days, reported separately, corresponding to next-day gate operations, weekly scheduling, and two-week operational planning.

The evaluation rests on a single principle: a variable enters a model, or conditions a forecast, only if it is known at the time t at which the forecast is issued. Every design choice below follows from this principle.

The five-year record is partitioned once. The period 2015–2017 is used for model development and for internal selection through rolling-origin validation. The year 2018 is held out as the test period and is the source of the reported accuracy figures. The year 2019 is reserved as a temporal stress-test (Section 4) and is used for neither training nor selection. No test-period observation participates in model selection, look-back tuning, or component weighting.

Accuracy is estimated with a rolling origin that advances one day at a time across the evaluation period. At each origin t, every model receives only the information available up to and including t and issues forecasts for the targets $t + 1$, $t + 7$, and $t + 14$. No observation posterior to t is used as an input or for forecast construction. Every model thus solves

the same forecasting task, at the same horizons, under the same information constraints (Figure 1). Model-specific estimation and updating procedures are described in Section 3.5.

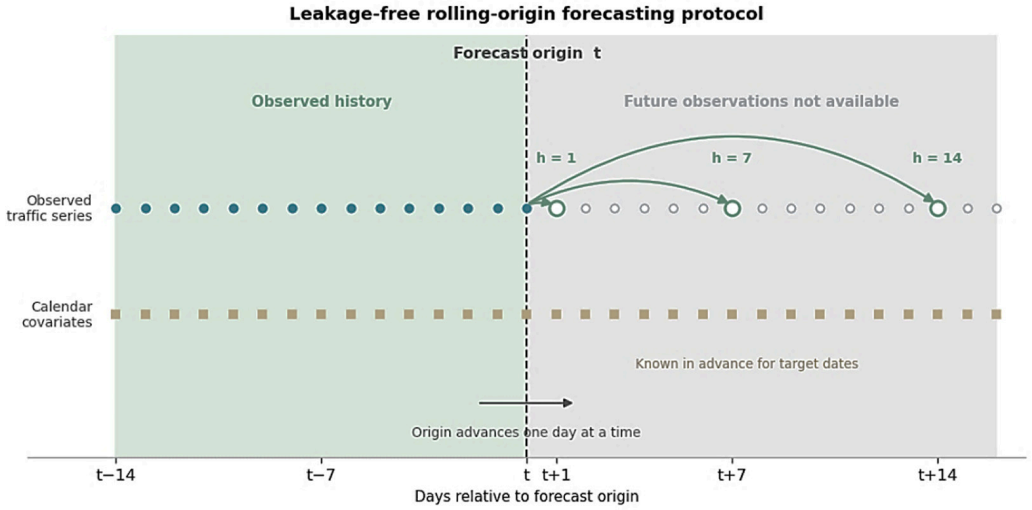


Figure 1. Leakage-free rolling-origin forecasting protocol. Circles represent traffic observations, and squares represent deterministic calendar covariates. At each forecast origin t, models use only observations available up to t and deterministic calendar attributes known for the target dates. Forecasts are evaluated at $h = 1, 7$, and 14 days as the origin advances one day at a time.

3.2. Data and Cleaning

The dataset was provided by the port authority and records vehicle access to the Port of Trois-Rivières over 2015–2019. The port operates two access gates with multiple entry and exit lanes. Each record carries a vehicle type (car or freight truck) and entry and exit timestamps. Because the prediction target is a count of arrivals, cleaning is based on arrival integrity.

A record is retained if and only if its entry timestamp and vehicle type are present, its entry timestamp falls within 2015–2019, it is not a strict duplicate (all fields identical, including a millisecond-resolution timestamp, which identifies a genuine logging duplicate), and its time in system is non-negative (rejecting the sole internal inconsistency, an exit recorded before entry). This procedure retains 775,983 records and yields two daily count series, cars and trucks, of 1826 days each (1 January 2015 to 31 December 2019), with mean daily volumes of approximately 128 trucks and 297 cars. Figure 2 shows the resulting series together with the development, test, and temporal stress-test periods defined in Section 3.1. A small number of truck records exhibit unusually long durations between entry and exit. Because they represent a marginal share of the data and the forecasting target is defined by the entry timestamp, records satisfying the arrival-integrity criteria above are retained in the analysis. Standard calendar attributes (day of week, month, weekday–weekend indicator, and public-holiday status) are derived from the calendar date associated with each observation and, for forecasting, from the known target date.

Visual inspection of the daily series also reveals pronounced periods of unusually low recorded traffic in 2019, suggesting possible disruptions or incompleteness in the gate records during parts of that year. Given this departure from the preceding observation period, 2019 was excluded from model development and selection and reserved as a temporal stress-test rather than treated as a conventional test year.

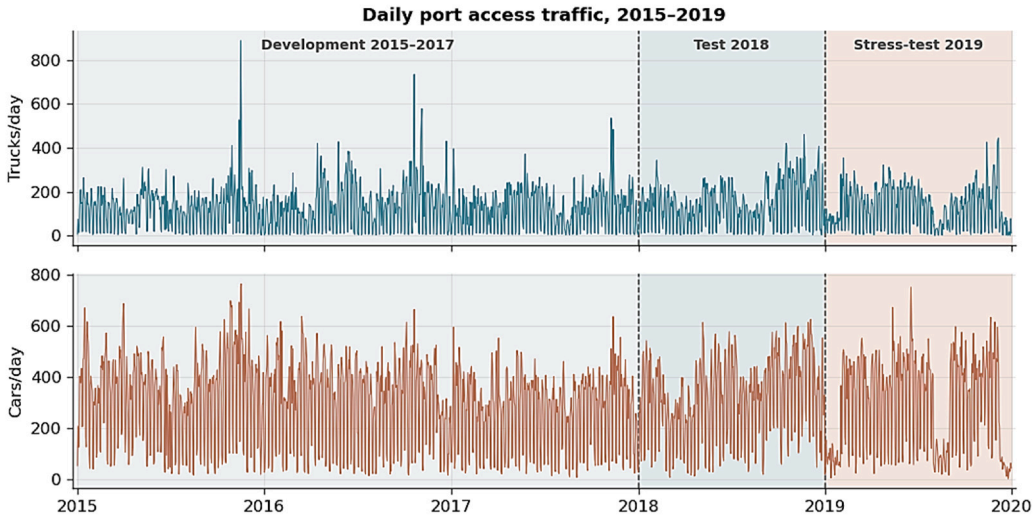


Figure 2. Daily truck and car arrivals at the port, 2015–2019. Shaded areas indicate the development (2015–2017), test (2018), and temporal stress-test (2019) periods.

3.3. Exploratory Analysis and Time-Series Diagnostics

Exploratory analysis on the development period characterizes the temporal structure of the series and motivates the modeling choices. The day-of-week profiles reveal a pronounced weekday–weekend contrast for both vehicle types, with relatively high weekday volumes followed by a sharp decline on Saturday and Sunday (Figure 3).

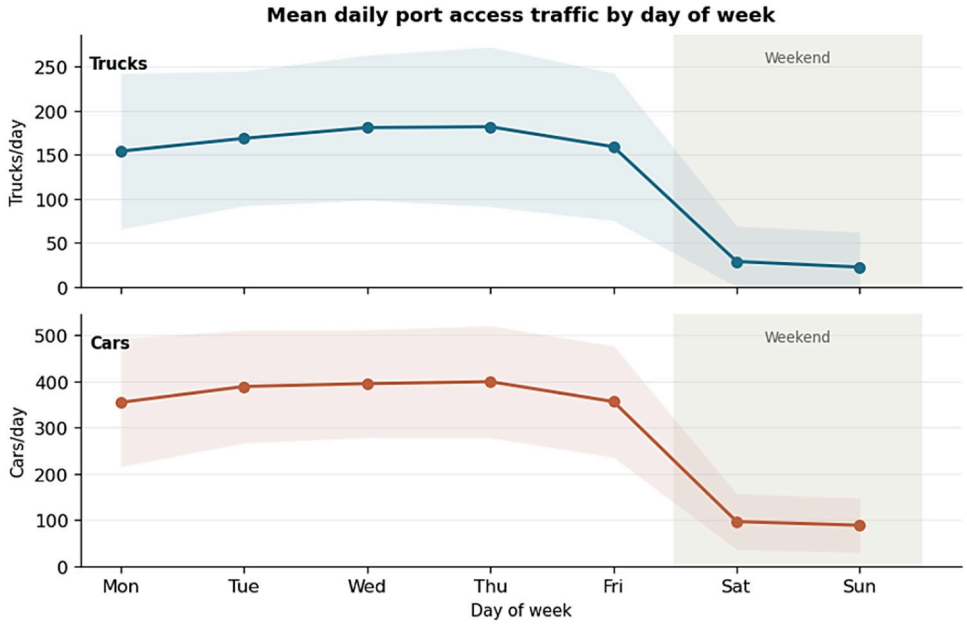


Figure 3. Mean daily port access traffic by day of week for trucks and cars. The shaded area identifies weekend days.

Time-series diagnostics further characterize this temporal structure (Figure 4). STL decomposition reveals a pronounced seven-day seasonal component in both series, while the autocorrelation functions exhibit recurring peaks at weekly lags. The partial autocorrelation functions additionally show substantial dependence at short lags.

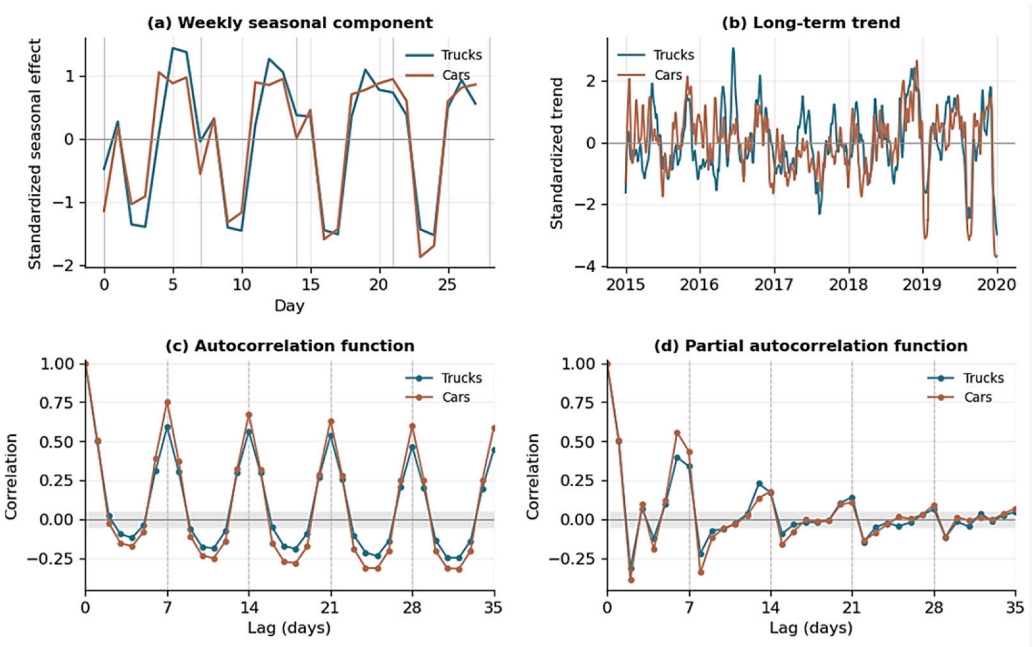


Figure 4. Time-series diagnostics for daily port access traffic: (**a**) standardized STL weekly seasonal component, (**b**) standardized STL trend component, (**c**) autocorrelation function, and (**d**) partial autocorrelation function for trucks and cars.

These diagnostics support three modeling decisions: strong weekly seasonality and structured autocorrelation justify the inclusion of a seasonal statistical model; irregular fluctuations beyond the recurring seasonal structure motivate the inclusion of nonlinear learners; and the coexistence of regular seasonality with irregular peaks motivates comparative evaluation across model families. The weekday–weekend structure is treated as an observable calendar regime available at the forecast origin. Whether regime-specific model assignment is warranted is an empirical question addressed in the comparative evaluation, not a design assumption.

3.4. Information Set and Feature Construction

For predicting the target at $t + h$ from origin t, the admissible information consists of observed target values up to t (genuine lags), deterministic calendar attributes of the target date (day of week, cyclic month, public holiday, weekend indicator), and any exogenous variable available at t. Calendar attributes are one-hot or cyclically encoded to avoid imposing a spurious ordinal scale on the day of week.

The meteorological variables examined here show only weak contemporaneous linear associations with daily traffic, with absolute Pearson correlations not exceeding 0.06 for either vehicle type (Figure 5). Given the absence of a meaningful contemporaneous linear signal in these data, they are not retained in the forecasting specification.

Car traffic is not used as a predictor of truck traffic: its association with truck volume is contemporaneous (a same-day partial correlation of 0.408, net of calendar) and therefore unavailable at the origin, while its lagged partial correlation is uninformative (-0.045). Calendar features are therefore retained to represent the observable temporal structure available at the forecast origin.

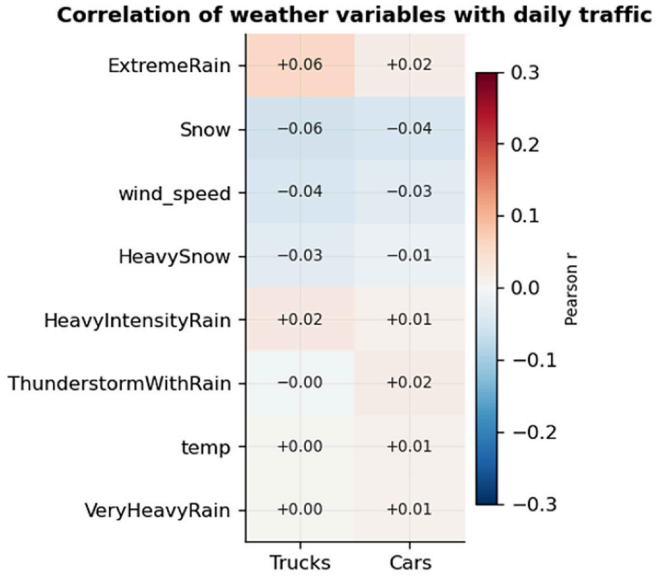


Figure 5. Pearson correlations between the examined meteorological variables and daily truck and car traffic.

3.5. Models and Comparative Evaluation

Under the common evaluation protocol defined in Section 3.1, the comparison includes two naïve baselines and four forecasting models. The two naïve baselines set a performance floor that any model must beat to be credited: a previous-day baseline and a seasonal (previous-week) baseline. The statistical model is SARIMAX(1,0,1)(1,1,1,7), with weekly seasonality identified from the autocorrelation diagnostics and orders selected by information criteria on the development period. The learning-based models are a Gradient Boosting Regressor (GBR), trained per horizon on lagged counts and target-date calendar features, and two recurrent networks, a GRU and an LSTM network, each reading a 14-day input window together with the target-date calendar features. The recurrent architecture uses a single recurrent layer of 64 units and a 14-day look-back window. Model configuration and look-back selection were conducted exclusively within the development period, before evaluation on the held-out 2018 test period. Neural training is identical across vehicle types, and predictions are averaged over three seeds; all stochastic elements are seeded and logged.

Temporal heterogeneity can be incorporated into forecasting in two conceptually distinct ways, and both are compared here. The first is regime conditioning: a single unified forecasting configuration is applied across all observations, while calendar information describing the target date is supplied among the predictors. This allows the forecasting function to adapt to weekday–weekend differences without changing the forecasting configuration across regimes. In this study, the unified conditioned configuration combines the GRU and GBR using weights fixed exclusively from the development period before test evaluation. The GRU weight is 0.9 for trucks and 0.7 for cars, with complementary GBR weights of 0.1 and 0.3, respectively; these weights are kept unchanged for the 2018 test and 2019 stress-test periods.

The second strategy is regime-based model assignment: distinct forecasting configurations are assigned to predefined temporal regimes, motivated by the hypothesis that structurally different regimes may favor models with different inductive biases. Based on the contrasting temporal characteristics identified during development (Section 3.3), a candidate fixed assignment rule was specified independently of the 2018 test period: the GRU–GBR ensemble was applied to weekday targets and SARIMAX to weekend targets. The weekend regime exhibits lower volume and greater regularity (Section 3.3), characteris-

tics that motivated the use of a model with explicit seasonal structure, whereas the greater variability of weekday traffic motivated the use of the flexible non-linear ensemble.

This fixed assignment constituted a candidate specialization rule specified before evaluation on the held-out 2018 test period; it was not defined as the empirical winner for every vehicle type, regime, and forecast horizon in the development data. Its empirical support was assessed through regime-specific performance within the development period and subsequently evaluated without modification on the held-out 2018 test period; no test-period information was used to define or modify the rule. Its empirical support was assessed through regime-specific performance within the development period and subsequently evaluated without modification on the held-out 2018 test period; no test-period information was used to define or modify the rule. This design assesses not only whether regime-based assignment improves forecast accuracy, but also whether the relative superiority of the assigned models is sufficiently stable across forecast horizons and evaluation periods to justify a fixed assignment rule. Both strategies are evaluated under the same leakage-free rolling-origin protocol and at the same horizons, $h \in \{1, 7, 14\}$, holding the information set, forecast origins, horizons, and evaluation periods constant. For each target date, the weekday–weekend regime is determined from calendar information known at forecast origin t. The fixed assignment involves no online regime detection or look-ahead.

Accuracy is measured by mean absolute error, root mean squared error, and the coefficient of determination, computed per horizon and per period and pooled over the rolling origins. The same metrics are reported per regime (weekday and weekend, by vehicle type). Statistical significance of forecast-accuracy differences is assessed with the Diebold–Mariano test, computed per horizon under a quadratic loss with a small-sample correction; residual autocorrelation is assessed with the Ljung–Box test; and the Wilcoxon signed-rank test on paired absolute errors is used where residual structure warrants it. Prediction intervals, where shown, are formed from empirical quantiles of the evaluation residuals and are labelled as uncalibrated; interval calibration is beyond the scope of this study.

4. Results

4.1. Comparative Predictive Performance on the 2018 Test Period

All models were evaluated under the rolling-origin protocol of Section 3 on the held-out 2018 test period at horizons $h \in \{1, 7, 14\}$. Table 2 reports the pooled predictive performance. At the one-day horizon, the previous-day baseline yields low explanatory power (R^2 = 0.068 for trucks and 0.015 for cars), whereas the seasonal baseline performs substantially better (R^2 = 0.349 and 0.482, respectively), consistent with the strong weekly structure identified in Section 3.3. Among the individual models, no single model dominates across vehicle types and horizons. At the one-day horizon, GRU provides the highest R^2 among the individual models for both trucks and cars, while relative performance becomes closer and more model-dependent at the longer horizons. The unified GRU–GBR ensemble attains the highest or near-highest pooled performance among the non-assignment configurations, with R^2 = 0.748 for trucks and 0.796 for cars at the primary one-day horizon.

On the 2018 test period, the fixed regime assignment attains the highest car R^2 at every horizon (0.823, 0.688, 0.660), improving on the unified conditioned ensemble by up to 0.03 (Diebold–Mariano $p \leq 0.001$); for trucks, the two configurations are statistically indistinguishable ($p \geq 0.21$). Read on the test year alone, the assignment is therefore the more accurate configuration for cars and no worse for trucks, and we report this result before turning to how it should be interpreted.

Because the assignment was prespecified rather than selected from 2018 performance, evaluating its advantage requires examining the corresponding model ordering on the development record. On the 2017 validation period, the assignment shows no correspond-

ing advantage over the unified conditioned ensemble: at the one-day operational horizon it is marginally lower for both vehicle types (cars 0.792 vs. 0.797, trucks 0.670 vs. 0.688). Whether the 2018 advantage reflects a consistent model–regime ordering across periods is examined in Sections 4.2 and 4.3. Both composite configurations beat the seasonal baseline at every horizon ($p < 0.001$).

Table 2. Test-2018 R^2 (pooled over rolling origins) by model, vehicle, and horizon.

Model	Truck $h = 1$	Truck $h = 7$	Truck $h = 14$	Car $h = 1$	Car $h = 7$	Car $h = 14$
Previous-day	0.068	0.344	0.307	0.015	0.474	0.366
Seasonal (−7 d)	0.349	0.344	0.307	0.482	0.474	0.366
SARIMAX	0.684	0.514	0.474	0.773	0.623	0.585
GBR	0.692	0.557	0.575	0.749	0.601	0.578
LSTM	0.744	0.633	0.636	0.773	0.668	0.649
GRU	0.746	0.631	0.637	0.787	0.670	0.655
Unified ensemble	0.748	0.634	0.638	0.796	0.667	0.646

Figure 6 provides a model-level view of these results across the three forecast horizons. The recurrent and composite models perform particularly strongly at the one-day horizon, while the relative differences among models change as the forecast horizon increases.

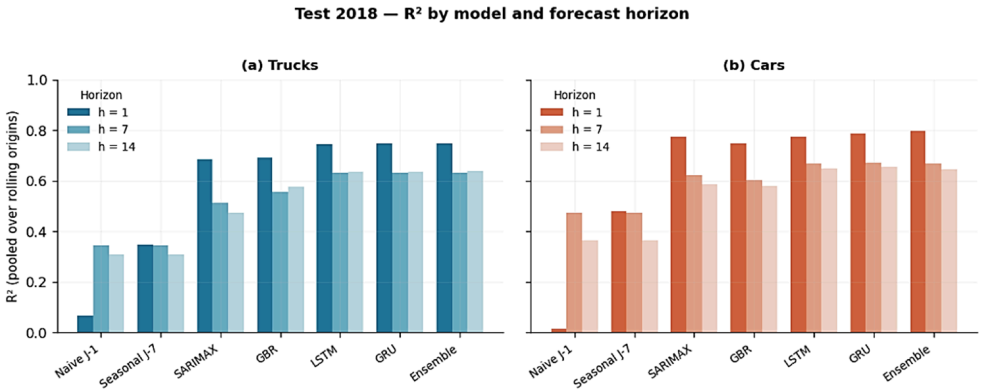


Figure 6. R^2 by forecasting model and horizon on the 2018 test period for (**a**) trucks and (**b**) cars.

For illustration, Figure 7 compares the one-day-ahead unified ensemble forecasts with observed truck and car arrivals during March 2018. The forecasts reproduce the recurring short-term fluctuations in both series, including the weekday–weekend transitions, while deviations remain around individual daily peaks and troughs.

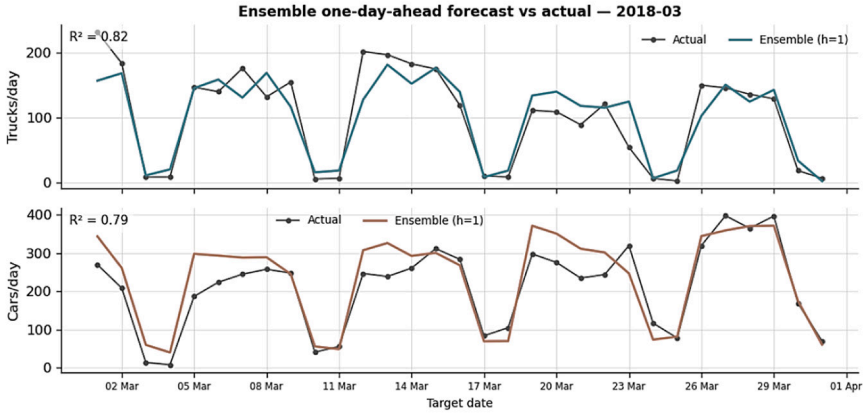


Figure 7. One-day-ahead unified ensemble forecasts and observed arrivals during March 2018 for trucks and cars.

4.2. Whether the Test-Period Advantage Is Supported Across Development and Test

On the 2018 held-out test period, the fixed regime-based assignment attains the highest car R^2 at every horizon (0.823, 0.688, and 0.660), improving on the unified conditioned ensemble by up to 0.03 (Diebold–Mariano $p \leq 0.001$). For trucks, the two configurations are statistically indistinguishable ($p \geq 0.21$). Considered on the 2018 test period alone, the assignment is therefore more accurate for cars and no worse for trucks. Because the assignment was fixed before evaluation on the 2018 test period and was not selected from 2018 performance, its apparent advantage must also be considered against its behavior within the development record. On the 2017 validation period, the assignment shows no corresponding advantage over the unified conditioned ensemble: at the primary one-day horizon, its R^2 is marginally lower for both cars (0.792 vs. 0.797) and trucks (0.670 vs. 0.688). Whether this difference reflects sufficiently stable model–regime dominance across periods and horizons is examined in Section 4.3.

The validation-to-test behavior of the unified ensemble is summarized in Figure 8. For trucks, R^2 increases from validation to test at all three horizons, from 0.688 to 0.748 at $h = 1$, from 0.593 to 0.634 at h = 7, and from 0.595 to 0.638 at h = 14. For cars, R^2 changes only marginally at the primary one-day horizon (0.797 to 0.796), while larger validation-to-test decreases occur at the one- and two-week horizons. Overall, the unified ensemble therefore maintains comparable or improved performance at the primary operational horizon, while generalization becomes more horizon-dependent at the longer horizons.

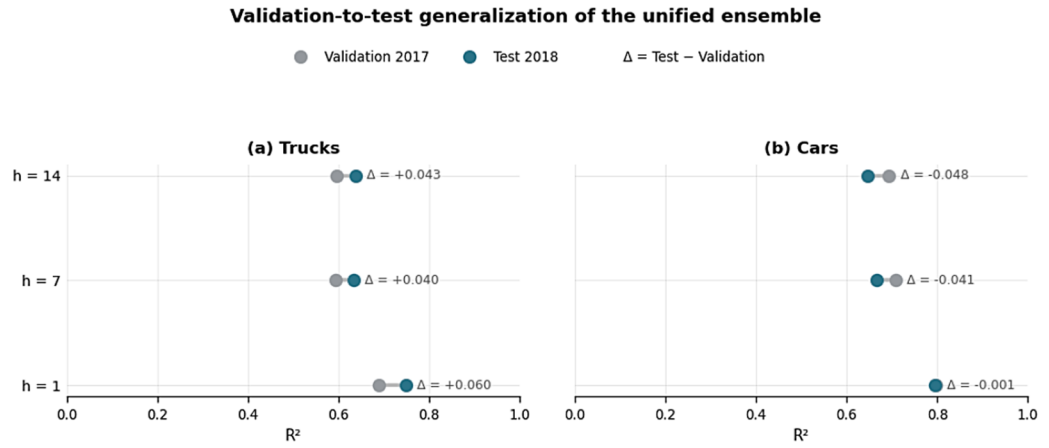


Figure 8. Validation-to-test generalization of the unified ensemble for (**a**) trucks and (**b**) cars at $h = 1$, 7, and 14 days. Points report R^2 for the 2017 validation and 2018 test periods.

The development record shows that the weekend model ordering varies across horizons and differs between evaluation periods. On the weekend car regime at the one-day operational horizon, the 2017 validation ordering is reversed relative to 2018: the neural ensemble is more accurate than SARIMAX (mean absolute error 29.74 vs. 35.90), whereas on the 2018 test SARIMAX leads (43.28 vs. 56.50). At the longer horizons SARIMAX becomes slightly more accurate on the 2017 weekend regime (35.60 vs. 35.04 at $h = 7$; 36.36 vs. 35.29 at $h = 14$), whereas its advantage is substantially larger across all horizons in 2018 (68.24 vs. 58.66 at $h = 7$; 68.11 vs. 60.95 at $h = 14$). The same horizon pattern holds for trucks, where the neural ensemble is the more accurate weekend configuration at $h = 1$ on both periods.

Consequently, the test-period advantage of fixed assignment was not supported by consistent model–regime dominance across development and test periods. The apparent 2018 advantage of the fixed assignment coincides with a reversal in regime-specific model dominance between validation and test, rather than with a model–regime relationship consistently observed during development.

4.3. Behavior Across Temporal Regimes

The fixed regime-based assignment evaluated against the unified conditioned configuration assigns the GRU–GBR ensemble to weekday targets and SARIMAX to weekend targets, as illustrated in Figure 9.

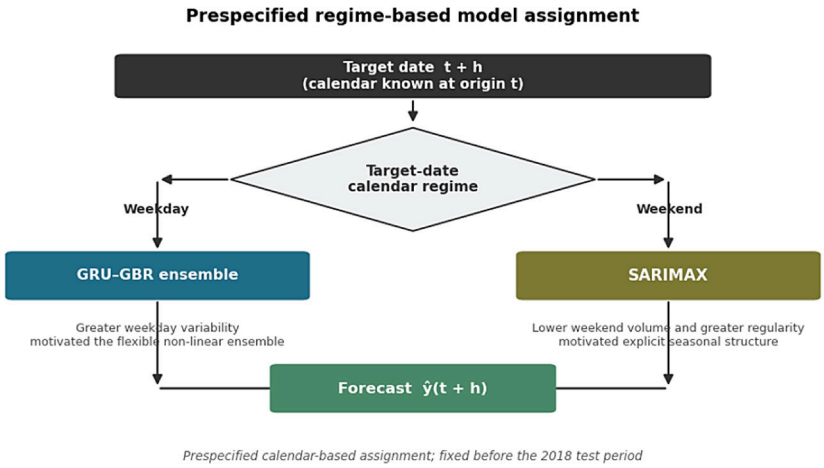


Figure 9. Prespecified regime-based model assignment. The target-date calendar regime, known at forecast origin t, assigns the GRU–GBR ensemble to weekday targets and SARIMAX to weekend targets.

Table 3 examines the relative performance of the GRU–GBR ensemble and SARIMAX within each calendar regime and forecast horizon, first on the 2017 validation period (Panel A) and then on the 2018 held-out test period (Panel B). Errors are reported as mean absolute error (MAE), because the coefficient of determination computed within a single regime becomes less informative at the longer horizons as the target variance is reduced.

The weekday side of the rule is consistently supported across the reported comparisons: the GRU-GBR ensemble is more accurate than SARIMAX on weekday targets for both vehicles, at every horizon, in both the validation and the test period. Weekend model ordering is more horizon- and period-dependent. For trucks, the GRU–GBR ensemble is more accurate at the one-day horizon in both periods (MAE 19.94 vs. 25.75 in 2017; 21.46 vs. 24.83 in 2018), whereas SARIMAX is more accurate at the 7- and 14-day horizons. For cars, the most pronounced instability occurs at the one-day horizon: the GRU–GBR ensemble is more accurate on the 2017 validation period (29.74 vs. 35.90), whereas SARIMAX is more accurate on the 2018 test period (43.28 vs. 56.50). At the longer horizons, SARIMAX is slightly more accurate for weekend cars in 2017 (h = 7: 35.04 vs. 35.60; h = 14: 35.29 vs. 36.36), and its advantage becomes substantially larger in 2018 (h = 7: 58.66 vs. 68.24; h = 14: 60.95 vs. 68.11).

Consequently, the test-period advantage of fixed assignment is not supported by consistent model–regime dominance across development and test periods. Although the weekday side of the assignment is consistently supported, the weekend side is horizon-dependent and, for cars at the one-day horizon, reverses between validation and test. The apparent 2018 advantage of fixed assignment therefore does not correspond to a model–regime relationship consistently observed across the development and test periods.

Figure 10 provides a compact representation of these regime-specific comparisons. Weekday comparisons favor the GRU–GBR ensemble for both vehicle types (Figure 10a,c), whereas weekend comparisons vary with forecast horizon (Figure 10b,d). For cars at h = 1, the model ordering reverses between validation and test (Figure 10d), with the ensemble favored in 2017 and SARIMAX favored in 2018.

Table 3. Per-regime forecast error (MAE), ensemble vs. SARIMAX, by vehicle, regime, and horizon. The regime is the calendar type of the target date. Lowest error in bold. Panel A: 2017 validation. Panel B: 2018 test.

Panel	Vehicle	Regime	*h*	Ensemble	SARIMAX	Winner
Panel A—Validation 2017	Truck	Weekday	1	**41.25**	46.73	Ensemble
	Truck	Weekday	7	**49.08**	55.63	Ensemble
	Truck	Weekday	14	**48.04**	57.05	Ensemble
	Truck	Weekend	1	**19.94**	25.75	Ensemble
	Truck	Weekend	7	23.05	**18.63**	SARIMAX
	Truck	Weekend	14	21.37	**18.47**	SARIMAX
	Car	Weekday	1	**55.75**	57.12	Ensemble
	Car	Weekday	7	**70.18**	71.92	Ensemble
	Car	Weekday	14	**71.39**	76.97	Ensemble
	Car	Weekend	1	**29.74**	35.90	Ensemble
	Car	Weekend	7	35.60	**35.04**	SARIMAX
	Car	Weekend	14	36.36	**35.29**	SARIMAX
Panel B—Test 2018	Truck	Weekday	1	**41.12**	48.21	Ensemble
	Truck	Weekday	7	**50.53**	61.65	Ensemble
	Truck	Weekday	14	**50.21**	65.02	Ensemble
	Truck	Weekend	1	**21.46**	24.83	Ensemble
	Truck	Weekend	7	26.36	**22.80**	SARIMAX
	Truck	Weekend	14	27.08	**24.21**	SARIMAX
	Car	Weekday	1	**56.84**	64.61	Ensemble
	Car	Weekday	7	**78.62**	85.84	Ensemble
	Car	Weekday	14	**80.82**	90.30	Ensemble
	Car	Weekend	1	56.50	**43.28**	SARIMAX
	Car	Weekend	7	68.24	**58.66**	SARIMAX
	Car	Weekend	14	68.11	**60.95**	SARIMAX

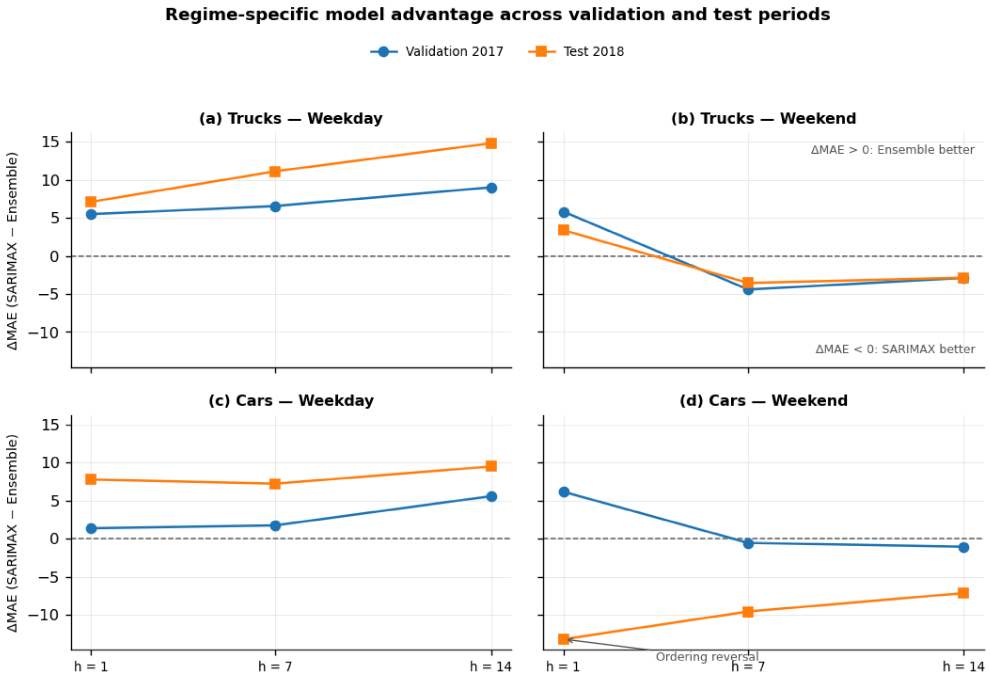


Figure 10. Regime-specific MAE difference between SARIMAX and the GRU–GBR ensemble for (**a**) trucks–weekday, (**b**) trucks–weekend, (**c**) cars–weekday, and (**d**) cars–weekend across forecast horizons in the 2017 validation and 2018 test periods. Positive values favor the ensemble; negative values favor SARIMAX.

Overall, the regime-specific results show that the weekday component of the fixed assignment is consistently supported across vehicles, horizons, and evaluation periods, whereas the weekend component is horizon-dependent and, for cars at the one-day horizon, reverses between validation and test. On this empirical basis, the unified calendar-conditioned ensemble is retained as the forecasting configuration evaluated in the subsequent residual diagnostics and temporal stress-test.

4.4. Residual Diagnostics

Residual autocorrelation is assessed with the Ljung–Box test on the retained unified ensemble's one-day-ahead residuals over the 2018 test period. For trucks, the test does not reject the null of no residual autocorrelation at either lag ($p = 0.08$ at lag 7; $p = 0.28$ at lag 14). For cars, the null is rejected at both the weekly and two-week lags ($p < 0.001$), indicating statistically detectable residual autocorrelation.

Figure 11 complements the Ljung–Box tests with the residual distributions and autocorrelation functions of the retained unified ensemble at h = 1. Truck residuals show limited remaining serial dependence (Figure 11a,c), whereas car residuals retain visible autocorrelation at weekly lags (Figure 11b,d), consistent with the corresponding Ljung–Box results.

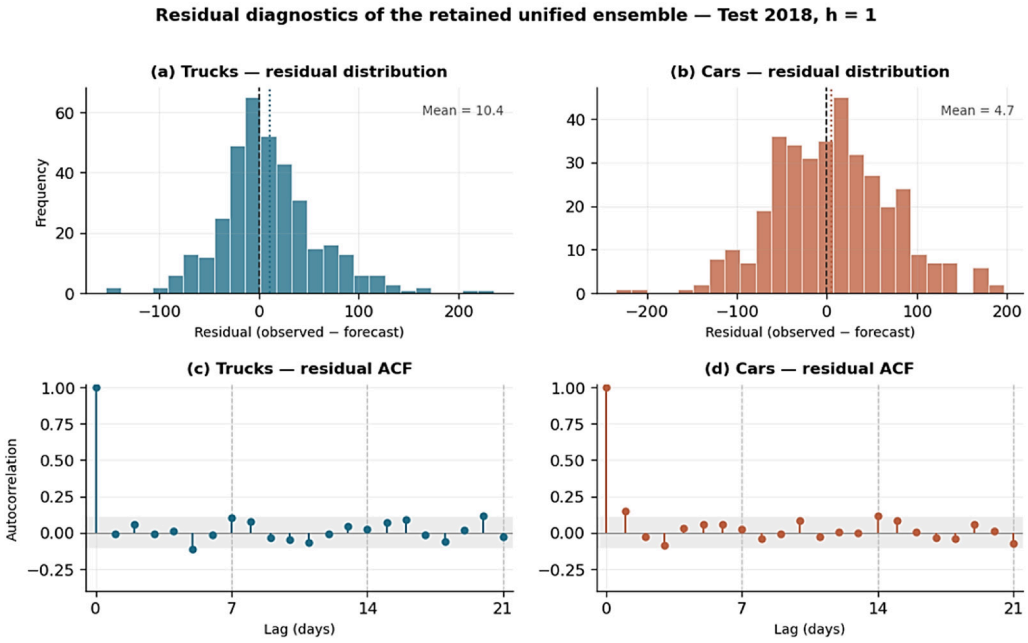


Figure 11. Residual diagnostics for the retained unified ensemble on the 2018 test period at $h = 1$: residual distributions for (**a**) trucks and (**b**) cars, and residual autocorrelation functions for (**c**) trucks and (**d**) cars. Dotted lines in (**c**,**d**) indicate the 95% confidence bounds for the autocorrelation function.

For comparison, the fixed regime-based assignment produces Ljung–Box p-values of 0.040 at lag 7 and 0.055 at lag 14 for cars on the 2018 test period. Residual autocorrelation therefore remains statistically detectable at lag 7, whereas the null is not rejected at lag 14 at the 5% significance level. This diagnostic indicates that the retained unified ensemble does not fully remove the weekly dependence in car residuals, which remains a limitation of the selected configuration.

4.5. Temporal Stress-Test on 2019

The year 2019, reserved from training and selection, serves as a temporal stress-test of the retained unified ensemble. At the one-day operational horizon, the coefficient of determination of the retained unified ensemble declines only modestly from 2018 to 2019 (trucks 0.748 → 0.734; cars 0.796 → 0.754). The decline becomes progressively larger at

the longer horizons (at one week, 0.634 → 0.556 for trucks and 0.667 → 0.559 for cars; at two weeks, 0.638 → 0.528 and 0.646 → 0.485), with larger reductions for cars than for trucks.

The largest forecast errors during the 2019 stress-test are concentrated in a limited number of months. At the one-day horizon, monthly MAE is highest in December, February, and November for trucks (Figure 12a) and in December, August, and January for cars (Figure 12b). Figure 12 shows the temporal distribution of these errors throughout the 2019 stress-test. Characterized from the data alone, 2019 exhibits localized departures from the traffic patterns observed during 2015–2018, particularly in the winter months and, for cars, in August, while broadly preserving the weekday–weekend structure. The causes of these departures cannot be established from the available data; 2019 is therefore treated as a temporal stress-test characterized by atypical traffic conditions rather than as a documented operational disruption.

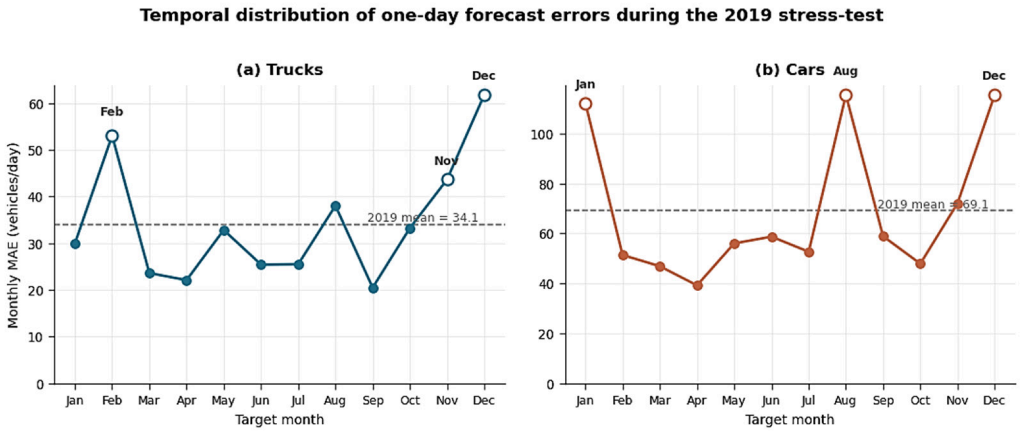


Figure 12. Monthly MAE of the retained unified ensemble during the 2019 temporal stress-test at h = 1 for (**a**) trucks and (**b**) cars. Dashed lines indicate the corresponding 2019 mean MAE.

The implications of the observed performance changes across forecast horizons are discussed in Section 5.3.

4.6. Descriptive Hourly Disaggregation

A daily forecast can be descriptively disaggregated into an hourly profile by distributing the predicted daily total according to historical empirical distributions by hour of day. This descriptive allocation illustrates how a daily forecast can be translated into an intra-day traffic profile; it does not constitute an independently estimated hourly forecast and remains complementary to the daily forecasting task evaluated in this study.

5. Discussion

5.1. Temporal Regimes, Model Dominance, and the Limits of Fixed Assignment

The exploratory analysis identifies a pronounced weekday–weekend heterogeneity in both volume and regularity for both vehicle types (Section 3.3). This observable structure provides a plausible basis for model specialization, but does not establish that different forecasting models should be assigned to the two regimes. The comparative results therefore distinguish the existence of temporal heterogeneity from the stability of the model–regime relationships used to exploit it.

The results show that temporal heterogeneity is not equivalent to stable model–regime dominance. Model preference depends jointly on temporal regime, forecast horizon, and evaluation period. Most importantly, the same observable regime does not consistently imply the same preferred model: weekend model ordering varies with horizon and, for

cars at the one-day horizon, reverses between validation and test (Section 4.3). Regime membership alone is therefore insufficient to justify a fixed model assignment.

This yields a condition sharper than temporal heterogeneity for determining when fixed regime-based assignment is warranted. Fixed regime assignment requires not only identifiable temporal heterogeneity, but sufficiently stable model–regime dominance across the forecast horizons and evaluation periods relevant to its use. When relative model accuracy is itself horizon-dependent or temporally unstable, a fixed switching rule cannot be justified from the development evidence merely because it performs well on a particular held-out period. Regime-specific predictive superiority must therefore be demonstrated as sufficiently stable, rather than inferred from the existence of distinct temporal regimes.

The unified conditioned ensemble is retained on this basis rather than on a claim of uniform predictive superiority. The fixed assignment achieves higher car accuracy on the 2018 test period, but the model–regime dominance required to justify that assignment is not sufficiently stable across horizons and evaluation periods. The unified configuration exploits the same ex ante calendar information without committing to a fixed model–regime correspondence. Its retention therefore reflects the stability criterion established above, rather than superiority on every individual performance metric. This retention should not be interpreted as evidence that conditioning is universally superior to assignment; rather, it is the more defensible strategy for the present problem under the stability criterion examined here.

5.2. Regime- and Horizon-Dependent Model Behavior

The results also show that model behavior cannot be characterized independently of forecast horizon. At the one-day horizon, the GRU–GBR ensemble is particularly competitive, including on weekend observations. At the 7- and 14-day horizons, however, SARIMAX becomes more competitive on weekend targets, while the ensemble retains its advantage on weekdays. The relative performance of statistical and nonlinear models therefore depends not only on the temporal regime being forecast, but also on how far ahead the forecast is issued.

This pattern is consistent with differences in the temporal information exploited by the model families. At short horizons, nonlinear learners can exploit recent observations and interactions among lagged and calendar predictors, whereas at longer horizons the recurring weekly structure captured explicitly by SARIMAX may become relatively more informative for weekend targets. The results do not establish these mechanisms causally, but they indicate that the usefulness of model specialization changes with forecast horizon.

These findings qualify the idea of regime-specific model specialization. A model that is preferable for a given temporal regime at one horizon need not remain preferable for that regime at another horizon. Forecast horizon should therefore be treated as an explicit dimension when evaluating regime-dependent forecasting strategies, rather than assuming a single model–regime correspondence across operational timescales.

5.3. Accuracy, Robustness, and Residual Structure

Assessing the retained configuration requires considering point accuracy, performance under temporal stress, and residual structure jointly. These dimensions do not all point in the same direction and therefore provide a more qualified assessment than any single performance metric.

On the 2018 held-out test period, the unified conditioned ensemble maintains strong one-day predictive performance, although fixed assignment achieves higher car accuracy (Section 4.2). Under the 2019 temporal stress-test, the decline remains modest at the primary one-day horizon but becomes progressively larger at the one- and two-week

horizons, particularly for cars (Section 4.5). The largest forecast errors are concentrated in a limited number of months rather than being uniformly distributed throughout the stress-test year.

Residual diagnostics provide a more qualified picture. For trucks, the Ljung–Box tests do not reject the null of no residual autocorrelation at lags 7 and 14, suggesting limited remaining serial dependence at the weekly and two-week scales examined. For cars, by contrast, the null is rejected at both lags, indicating statistically detectable residual autocorrelation. The fixed assignment produces a more favorable car residual diagnostic on the 2018 test period, although residual autocorrelation remains statistically detectable at lag 7. The evidence therefore does not uniformly favor one configuration on every criterion: fixed assignment provides higher car accuracy and a more favorable car residual diagnostic on the 2018 test period, whereas the unified configuration avoids dependence on a fixed model–regime correspondence whose stability is not supported across horizons and evaluation periods.

The combined evidence characterizes the retained configuration as robust primarily at the one-day operational horizon rather than uniformly across all horizons. It combines strong held-out accuracy with only modest degradation at that horizon under the 2019 temporal stress-test, while performance deteriorates more substantially at one and two weeks. This robustness is nevertheless incomplete: statistically detectable weekly dependence remains in the car residuals. That remaining structure identifies a concrete limitation of the retained configuration and motivates investigation of richer temporal or exogenous information, as discussed in Section 5.7.

5.4. Interpretability and Operational Relevance

The retained configuration is not intrinsically interpretable: the GRU–GBR ensemble remains a flexible, largely opaque forecasting architecture. Its information basis and forecasting protocol are nevertheless transparent and auditable. The information set is explicitly defined as observed arrival counts available up to the forecast origin together with deterministic calendar attributes of the target date, with no variable unavailable at the time of forecasting (Section 3.4). The forecasting process can therefore be audited against a documented admissible information set, even though the internal mapping learned by the recurrent component is not directly interpretable. The contribution should therefore be understood in terms of transparency and auditability of the information set and forecasting pipeline, rather than intrinsic interpretability of the ensemble itself.

Short- and medium-term forecasts of car and truck arrivals can support gate operations, resource allocation, and capacity planning. The present study evaluates the forecasting component under a leakage-free, out-of-sample protocol. Potential operational uses, such as aligning staffing and equipment with anticipated demand or coordinating responses to anticipated high-demand periods, remain contingent on organizational integration and decision processes; their realized operational benefits are not measured in this study.

5.5. Positioning Against Recent Deep-Learning Forecasting Literature

Recent forecasting research has increasingly adopted deep-learning architectures designed to capture complex temporal dependencies and long-range interactions, including attention-based, convolutional, and Transformer-type models, as reviewed in Section 2.4. These developments broaden the set of candidate forecasting architectures beyond the statistical, recurrent, and tree-based models evaluated here. The present study does not compare its GRU–GBR ensemble directly with these more recent architectures, and no claim of superiority over them is therefore made.

The contribution addressed here is instead methodological: given an observable temporal structure known at forecast origin, does predictive performance support conditioning a unified configuration on that information or assigning different models to different regimes? This question is not resolved by increasing model complexity alone. A more sophisticated architecture could improve forecast accuracy while leaving open whether regime-specific model dominance is sufficiently stable across horizons and evaluation periods to justify fixed assignment.

Recent deep-learning architectures therefore represent an important extension of the candidate model set rather than an alternative to the evaluation question examined in this study.

5.6. Transferability

The transferability of the results is stated conservatively. The empirical findings are specific to the Port of Trois-Rivières and to a five-year record ending in 2019, and the reported accuracy figures should not be assumed to hold for other facilities. The reusable element is methodological. Where an access-controlled system exhibits identifiable, calendar-driven temporal heterogeneity, the leakage-free rolling-origin protocol used here provides a template for evaluating whether that heterogeneity is better exploited by conditioning a unified forecasting configuration or by fixed regime-based assignment. A methodological implication follows from the present analysis: where fixed regime assignment is considered on the basis of regime-specific predictive superiority, that superiority should be tested for sufficient stability across the forecast horizons and evaluation periods relevant to deployment, rather than inferred from the mere existence of identifiable temporal regimes. Whether the specific configuration retained here generalizes to container terminals, airports, border crossings, or other logistics systems remains a hypothesis requiring external validation on independent data.

5.7. Limitations and Future Research

Several limitations frame the contribution. The analysis rests on a single facility over a bounded period; the quantitative results are context-specific, and generalization requires validation on independent datasets. The admissible information set is deliberately parsimonious, relying primarily on lagged traffic observations and deterministic calendar attributes. Richer exogenous information known at the forecast origin, such as vessel-call schedules and corridor-level industrial activity indices, could be examined as potential sources of information for addressing the weekly dependence that remains in the car residuals of the retained unified configuration (Section 4.4). The meteorological variables examined here showed little contemporaneous linear association with daily traffic and were therefore not retained in the admissible information set. The regime definition is static and calendar-based; online regime detection or rolling-window identification could be investigated as alternatives for adapting the regime representation to structural shifts. Prediction intervals are formed from empirical residual quantiles and are uncalibrated; interval calibration is left to future work. The candidate model set does not include recent Transformer-based forecasting architectures; benchmarking such models under the same leakage-free rolling-origin protocol and comparable model-selection budgets constitutes a further experimental extension.

6. Conclusions

This study examined short- and medium-term forecasting of daily car and freight-truck access traffic at an urban non-containerized port using a five-year operational record and a leakage-free rolling-origin evaluation of statistical, machine-learning, and composite

configurations. Its guiding question was whether identifiable calendar-based temporal heterogeneity is better exploited by conditioning a unified forecasting configuration or by assigning different models to different regimes.

The study makes three contributions. Scientifically, it shows that temporal heterogeneity is not equivalent to stable model–regime dominance: model preference varies with temporal regime and forecast horizon and can also change across evaluation periods. This yields a falsifiable condition for fixed regime-based forecasting: model assignment requires sufficiently stable model–regime dominance across the horizons and evaluation periods relevant to its use, a condition not supported by the present evidence. Methodologically, the study provides a leakage-free rolling-origin framework that aligns forecasts with their target dates, restricts predictors to information available at the forecast origin, and keeps development, test, and stress-test periods distinct. In applied terms, the retained unified calendar-conditioned ensemble maintains strong held-out performance and shows only modest degradation at the primary one-day operational horizon under the 2019 temporal stress-test, although performance deteriorates more substantially at the longer horizons and statistically detectable weekly dependence remains in the car residuals.

Future research should examine more adaptive regime definitions and richer exogenous information available at the forecast origin, formally calibrate predictive uncertainty, and benchmark recent forecasting architectures under the same leakage-free protocol. External validation on independent access-controlled systems is also required before claiming transferability beyond the present case. More broadly, the findings show that temporal heterogeneity can motivate regime-aware forecasting, but fixed model specialization should be retained only when the corresponding model–regime dominance is shown to be sufficiently stable rather than inferred from the existence of distinct temporal patterns.

Author Contributions: Conceptualization, B.B.-D.; Methodology, B.B.-D. and J.-F.A.; Software, B.B.-D.; Validation, B.B.-D. and J.-F.A.; Formal analysis, B.B.-D.; Investigation, B.B.-D.; Resources, B.B.-D.; Data curation, B.B.-D.; Writing—original draft, B.B.-D.; Writing—review & editing, J.-F.A.; Visualization, B.B.-D.; Supervision, J.-F.A.; Project administration, J.-F.A.; Funding acquisition, J.-F.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the PLAINE Program of the Réseau Québec Maritime and the CREATE Program of the Natural Sciences and Engineering Research Council of Canada (NSERC).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author due to confidentiality agreements with the collaborating port authority.

Acknowledgments: The authors acknowledge the collaboration with the Trois-Rivières Port Authority and the funding support from the PLAINE Program of the Réseau Québec Maritime, as well as the CREATE Program of the Natural Sciences and Engineering Research Council of Canada (NSERC).

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

References

1. Fan, D.; Heminthavong, K. *Road Transportation: Heavyweight of the Canadian Economy*; HillStudies Publication No. 2022-04-E; Library of Parliament: Ottawa, ON, Canada, 2022.
2. Lehmacher, W.; Lind, M.; Poikonen, J.; Meseguer, J.; Cárcel Cervera, J.L. Reducing Port City Congestion through Data Analysis, Simulation, and Artificial Intelligence to Improve the Well-Being of Citizens. *J. Mega Infrastruct. Sustain. Dev.* **2022**, *2*, 65–82. [CrossRef]
3. Ben Daya, B.; Audy, J.-F. Port Access Fluidity Management during a Major Extension Project: A Simulation-Based Case Study. *Sustainability* **2024**, *16*, 2834. [CrossRef]

4. Fleming, M.; Huynh, N.; Xie, Y. Agent-Based Simulation Tool for Evaluating Pooled Queue Performance at Marine Container Terminals. *Transp. Res. Rec.* **2013**, *2330*, 103–112. [CrossRef]
5. Guan, C.Q.; Liu, R. Modeling Gate Congestion of Marine Container Terminals, Truck Waiting Cost, and Optimization. *Transp. Res. Rec.* **2009**, *2100*, 58–67. [CrossRef]
6. Parola, F.; Satta, G.; Notteboom, T.; Persico, L. Revisiting Traffic Forecasting by Port Authorities in the Context of Port Planning and Development. *Marit. Econ. Logist.* **2021**, *23*, 444–494. [CrossRef]
7. Jahangard, M.; Xie, Y.; Feng, Y. Leveraging Machine Learning and Optimization Models for Enhanced Seaport Efficiency. *Marit. Econ. Logist.* **2025**, *27*, 710–751. [CrossRef]
8. Jin, J.; Ma, M.; Jin, H.; Cui, T.; Bai, R. Container Terminal Daily Gate In and Gate Out Forecasting Using Machine Learning Methods. *Transp. Policy* **2023**, *132*, 163–174. [CrossRef]
9. Ali Nadi, A.; Sharma, S.; Snelder, M.; Bakri, T.; van Lint, H.; Tavasszy, L. Short-Term Prediction of Outbound Truck Traffic from the Exchange of Information in Logistics Hubs: A Case Study for the Port of Rotterdam. *Transp. Res. Part C Emerg. Technol.* **2021**, *127*, 103111. [CrossRef]
10. Yu, H.; Cui, X.; Bai, X.; Chen, C.; Xu, L. Incorporating Graph Theory and Time Series Analysis for Fine-Grained Traffic Flow Prediction in Port Areas. *Ocean Eng.* **2025**, *335*, 121693. [CrossRef]
11. Mamede, F.P.; da Silva, R.F.; de Brito Junior, I.; Yoshizaki, H.T.Y.; Hino, C.M.; Cugnasca, C.E. Deep Learning and Statistical Models for Forecasting Transportation Demand: A Case Study of Multiple Distribution Centers. *Logistics* **2023**, *7*, 86. [CrossRef]
12. Orozco Torres, J.A.; Medina Santiago, A.; García-Martínez, J.R.; López-Zapata, B.Y.; Mijangos López, J.A.; Rincón Zapata, O.J.; Avitia López, J.A. A Data-Driven Approach Using Recurrent Neural Networks for Material Demand Forecasting in Manufacturing. *Logistics* **2025**, *9*, 130. [CrossRef]
13. Rashed, Y.; Meersman, H.; Van de Voorde, E.; Vanelslander, T. Short-Term Forecast of Container Throughput: An ARIMA-Intervention Model for the Port of Antwerp. *Marit. Econ. Logist.* **2017**, *19*, 749–764. [CrossRef]
14. Farhan, J.; Ong, G.P. Forecasting Seasonal Container Throughput at International Ports Using SARIMA Models. *Marit. Econ. Logist.* **2018**, *20*, 131–148. [CrossRef]
15. Alharbi, F.R.; Csala, D. A Seasonal Autoregressive Integrated Moving Average with Exogenous Factors (SARIMAX) Forecasting Model-Based Time Series Approach. *Inventions* **2022**, *7*, 94. [CrossRef]
16. Fiskin, C.S.; Turgut, O.; Westgaard, S.; Cerit, A.G. Time Series Forecasting of Domestic Shipping Market: Comparison of SARIMAX, ANN-Based Models and SARIMAX–ANN Hybrid Model. *Int. J. Shipp. Transp. Logist.* **2022**, *14*, 193–221. [CrossRef]
17. Yang, C.-H.; Chang, P.-Y. Forecasting the Demand for Container Throughput Using a Mixed-Precision Neural Architecture Based on CNN–LSTM. *Mathematics* **2020**, *8*, 1784. [CrossRef]
18. Wu, G.; Hu, Y.-C.; Chiu, Y.-J.; Jiang, P.; Chi, R. Forecast Combination Using Grey Relational Analysis and Choquet Fuzzy Integral for Container Throughput Forecasting. *Expert Syst. Appl.* **2024**, *252*, 124170. [CrossRef]
19. ArunKumar, K.E.; Kalaga, D.V.; Sai Kumar, C.M.; Kawaji, M.; Brenza, T.M. Comparative Analysis of GRU, LSTM, ARIMA, and SARIMA Models for Forecasting COVID-19 Trends. *Alex. Eng. J.* **2022**, *61*, 7585–7603. [CrossRef]
20. Peirano, R.; Kristjanpoller, W.; Minutolo, M.C. Forecasting Inflation in Latin American Countries Using a SARIMA–LSTM Combination. *Soft Comput.* **2021**, *25*, 10851–10862. [CrossRef]
21. Zhang, D.; Li, X.; Wan, C.; Man, J. A Novel Hybrid Deep-Learning Framework for Medium-Term Container Throughput Forecasting: An Application to China's Guangzhou, Qingdao and Shanghai Hub Ports. *Marit. Econ. Logist.* **2024**, *26*, 44–73. [CrossRef]
22. Tashman, L.J. Out-of-Sample Tests of Forecasting Accuracy: An Analysis and Review. *Int. J. Forecast.* **2000**, *16*, 437–450. [CrossRef]
23. Bergmeir, C.; Benítez, J.M. On the Use of Cross-Validation for Time Series Predictor Evaluation. *Inf. Sci.* **2012**, *191*, 192–213. [CrossRef]
24. Cerqueira, V.; Torgo, L.; Mozetič, I. Evaluating Time Series Forecasting Models: An Empirical Study on Performance Estimation Methods. *Mach. Learn.* **2020**, *109*, 1997–2028. [CrossRef]
25. Bergmeir, C.; Hyndman, R.J.; Koo, B. A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction. *Comput. Stat. Data Anal.* **2018**, *120*, 70–83. [CrossRef]
26. Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. Leakage in Data Mining: Formulation, Detection, and Avoidance. *ACM Trans. Knowl. Discov. Data* **2012**, *6*, 1–21. [CrossRef]
27. Hewamalage, H.; Ackermann, K.; Bergmeir, C. Forecast Evaluation for Data Scientists: Common Pitfalls and Best Practices. *Data Min. Knowl. Discov.* **2023**, *37*, 788–832. [CrossRef] [PubMed]
28. Diebold, F.X.; Mariano, R.S. Comparing Predictive Accuracy. *J. Bus. Econ. Stat.* **2002**, *20*, 134–144. [CrossRef]
29. Harvey, D.; Leybourne, S.; Newbold, P. Testing the Equality of Prediction Mean Squared Errors. *Int. J. Forecast.* **1997**, *13*, 281–291. [CrossRef]
30. Hamilton, J.D. A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* **1989**, *57*, 357–384. [CrossRef]

31. Carvalho, A.X.; Tanner, M.A. Mixtures-of-Experts of Autoregressive Time Series. *IEEE Trans. Neural Netw.* **2005**, *16*, 39–56. [CrossRef] [PubMed]
32. Bates, J.M.; Granger, C.W.J. The Combination of Forecasts. *Oper. Res. Q.* **1969**, *20*, 451–468. [CrossRef]
33. Timmermann, A. Forecast Combinations. In *Handbook of Economic Forecasting*; Elsevier: Amsterdam, The Netherlands, 2006; Volume 1, pp. 135–196. [CrossRef]
34. Taylor, S.J.; Letham, B. Forecasting at Scale. *Am. Stat.* **2018**, *72*, 37–45. [CrossRef]
35. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 11106–11115. [CrossRef]
36. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22419–22430.
37. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series Is Worth 64 Words: Long-Term Forecasting with Transformers. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. [CrossRef]
38. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 7–11 May 2024. [CrossRef]
39. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. [CrossRef]
40. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
41. Challu, C.; Olivares, K.G.; Oreshkin, B.N.; Ramirez, F.G.; Mergenthaler-Canseco, M.; Dubrawski, A. NHITS: Neural Hierarchical Interpolation for Time Series Forecasting. *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 6989–6997. [CrossRef]
42. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are Transformers Effective for Time Series Forecasting? *Proc. AAAI Conf. Artif. Intell.* **2023**, *37*, 11121–11128. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.