جامعة جدة
University of Jeddah

# Maritime Demand and Supply Forecasting Using ARIMA, LSTM, and XGBoost with Explainable Artificial Intelligence

Albandri Tami, Lama Alturki, Ghaidaa Elazhary

*Department of Data Science, University of Jeddah, Jeddah, Saudi Arabia*

**Abstract**

Maritime transportation underpins global trade and economic integration, making accurate demand forecasting essential for strategic planning and operational decision-making. Conventional statistical approaches, most notably the AutoRegressive Integrated Moving Average (ARIMA) model, have long been used for maritime demand forecasting due to their interpretability and strong theoretical foundations. However, despite their conceptual simplicity, ARIMA models remain highly effective when applied to maritime demand data characterized by limited sample sizes and relatively smooth long-term trends. In such settings, the robustness and stability of classical time-series models can outweigh the potential advantages of more complex learning-based approaches. [1, 2]

This study adopts a differentiated modeling strategy for maritime demand and supply forecasting. Due to differences in data availability and temporal coverage, maritime demand is modeled using ARIMA to ensure robustness under limited data conditions, while maritime supply—supported by a longer historical record—is analyzed using ARIMA, LSTM, and XGBoost to enable comparative evaluation. Explainable artificial intelligence (XAI) techniques are incorporated to interpret model predictions and highlight key contributing factors. [3, 4]

A strict time-based evaluation framework is adopted to ensure methodological rigor, with data from 1980 to 2020 used for training, and 2021 to 2025 reserved for testing. Model performance is assessed using standard regression metrics, including Percentage Error (PE) and Root Mean Squared Error (RMSE). The findings of this research provide empirical evidence supporting the suitability of ARIMA as the most reliable model for maritime demand forecasting, while advanced learning-based models are primarily evaluated in the context of maritime supply forecasting, where richer historical data enables meaningful comparative analysis. [5, 6]

*Keywords:* Maritime transportation, Demand forecasting, Time series analysis, ARIMA, LSTM, XGBoost, Explainable AI

## 1. Introduction

Maritime transportation is a cornerstone of the global economy, enabling the movement of raw materials, intermediate goods, and finished products across international markets. More

than eighty percent of global trade by volume is transported by sea, underscoring the strategic importance of maritime logistics in sustaining economic growth and supply chain resilience. As global trade patterns evolve due to economic cycles, geopolitical developments, and technological advancements, accurate forecasting of maritime demand has become increasingly critical. [2]

Reliable demand forecasts support a wide range of strategic and operational decisions, including fleet deployment, port capacity planning, infrastructure investment, and policy formulation. Inaccurate forecasts can lead to underutilized assets, congestion, financial losses, and inefficient resource allocation. Consequently, improving the accuracy and robustness of maritime demand forecasting models remains a central objective for researchers and industry practitioners alike.[5, 6]

Traditional approaches to maritime demand forecasting have largely relied on statistical time-series models. Among these, the Auto Regressive Integrated Moving Average (ARIMA) model has been widely adopted due to its mathematical rigor and ability to model linear temporal dependencies in univariate data. While ARIMA relies on linear assumptions, it has consistently demonstrated strong performance in maritime demand forecasting applications where data availability is limited and long-term trends remain relatively stable. In such contexts, its parsimonious structure and theoretical rigor contribute to robust and generalizable forecasts. [1]

Recent advances in machine learning and deep learning have introduced alternative modeling paradigms capable of capturing nonlinear dynamics and complex temporal interactions. These approaches are particularly relevant in maritime supply forecasting, where longer historical records and richer structural patterns allow learning-based models to demonstrate their strengths. Deep learning architectures, particularly Long Short-Term Memory (LSTM) networks, are designed to handle sequential data and retain long-term contextual information, making them suitable for time-series forecasting. In parallel, ensemble-based machine learning models such as Extreme Gradient Boosting (XGBoost) have gained prominence for their strong predictive performance, scalability, and ability to model nonlinear interactions among features. [3, 4]

Despite their predictive advantages, advanced learning-based models are often criticized for their lack of interpretability. In the context of maritime decision-making, stakeholders require not only accurate forecasts but also an understanding of the factors influencing these predictions. Explainable artificial intelligence (XAI) techniques have emerged to address this challenge by providing transparent explanations of model behavior and feature contributions. [7, 8]

This study evaluates maritime demand forecasting using ARIMA as a robust baseline model, reflecting its suitability under limited data conditions and stable long-term demand patterns. In contrast, advanced learning-based models, namely LSTM and XGBoost, are assessed exclusively for maritime supply forecasting, where extended historical data enables meaningful comparative evaluation of nonlinear modeling capabilities. By integrating XAI techniques to enhance interpretability, the research seeks to evaluate whether modern learning-based approaches can offer both improved accuracy and actionable insights for maritime decision support. It is important to note that the availability and temporal span of data differ between maritime supply and demand. While supply data is available over a long historical horizon, enabling the application and comparison of multiple forecasting models, demand data is more limited in length. Accordingly, this study emphasizes maritime supply as the primary focus of comparative model evaluation, while maritime demand forecasting is conducted using a stable baseline model. This design is further supported by empirical results, which confirm that ARIMA consistently outperforms more complex models in demand forecasting across all evaluated vessel categories. This modeling choice follows supervisory guidance and ensures methodological rigor under data availability

2

constraints.

## 2. Related Work

Table 1 summarizes representative studies related to maritime demand forecasting, supply-demand analysis, and machine learning applications in the maritime domain. The reviewed literature reflects a wide range of methodological approaches, including classical time-series models, econometric frameworks, and advanced learning-based techniques, with model performance varying depending on data availability, forecasting horizon, and application context.

Table 1: Summary of Related Work in Maritime Demand Forecasting

| Year | Study | Methodology | Data / Scope | Key Findings |
|---|---|---|---|---|
| 2023 | Huang et al. [5] | Demand prediction and sharing strategy | Maritime transportation demand under price and quality competition | Proposed a demand prediction framework considering resilience, pricing, and service quality in maritime transportation. |
| 2022 | Krikigianni et al. [6] | System design framework | Maritime supply chain demand–supply matching | Developed a framework to align emerging demand and supply needs in maritime supply chains. |
| 2022 | Martius et al. [3] | Machine learning forecasting | Worldwide empty container availability | Demonstrated that machine learning models improve forecasting accuracy for container availability. |
| 2022 | Mokhtar et al. [1] | ARIMA and SARIMA | Container throughput | Showed that ARIMA-based models are effective for short-term throughput forecasting. |
| 2023 | Munim et al. [4] | Prophet and hybrid time-series models | Major Asian ports | Hybrid models outperformed traditional time-series approaches in container throughput forecasting. |
| 2023 | Nomikos and Tsouknidis [2] | Econometric analysis | Shipping freight market demand and supply shocks | Identified distinct demand and supply shocks and their impact on shipping investments. |
| 2023 | Nourmohammadi et al. [8] | Deep spatiotemporal learning | Maritime accident prediction | Proposed a deep learning approach capturing spatiotemporal dependencies in maritime safety analysis. |
| 2023 | Park et al. [9] | Demand and supply shock analysis | Dry bulk shipping market | Analyzed the asymmetric impacts of demand and supply shocks on freight markets. |
| 2022 | Rong et al. [7] | Probabilistic prediction | Maritime traffic patterns | Introduced probabilistic prediction based on ship motion pattern extraction. |
| 2023 | Sui et al. [10] | Time series analysis | Maritime accidents in the Yangtze River | Applied time-series techniques to analyze accident trends and temporal patterns. |

Previous studies have examined maritime demand forecasting, supply–demand dynamics, and machine learning applications using a variety of statistical and data-driven approaches. While learning-based models have shown advantages in data-rich settings, classical time-series models such as ARIMA remain effective for demand forecasting under limited data conditions, supporting the modeling choices adopted in this study [5, 6, 3, 1, 4, 2, 8, 9, 7, 10].

## 3. Data Description

The dataset used in this study consists of historical maritime demand observations spanning from 1980 to 2020. The data represents aggregated annual measures of global maritime activity. The target variable captures maritime demand in numeric form, reflecting long-term trends in shipping activity over time.

Given the univariate nature of the demand forecasting task, the analysis relies primarily on historical demand values without extensive feature augmentation. [4]

## 4. Exploratory Data Analysis

Prior to model development, standard data cleaning procedures were applied to ensure data consistency and completeness. To preserve the temporal structure of the data and prevent information leakage, a time-based evaluation strategy was adopted, with historical observations used for model estimation and future periods reserved for out-of-sample forecasting. [4]

A focused exploratory data analysis (EDA) was conducted to examine the main statistical and temporal characteristics of the maritime dataset. The analysis provides high-level insights into long-term trends and variability patterns relevant to maritime demand and supply forecasting, without introducing unnecessary model complexity.

### 4.1. Descriptive Statistics

Table 2 Table 2 summarizes the descriptive statistics for the main vessel categories over the period 1980–2020. The results indicate substantial long-term growth in global fleet capacity, accompanied by varying levels of variability across vessel types. Container ships exhibit relatively higher variability, reflecting their rapid expansion over the study period, while other vessel categories demonstrate more stable growth patterns.

Table 2: Descriptive statistics of vessel categories (1980–2020, in million DWT)

| Vessel Type | Mean | Std Dev | Min | Max | Skewness | CV |
|---|---|---|---|---|---|---|
| Total Fleet | 850.2 | 420.5 | 420.3 | 1908.5 | 0.82 | 0.49 |
| Oil Tankers | 315.4 | 85.2 | 243.8 | 545.3 | 1.12 | 0.27 |
| Bulk Carriers | 301.2 | 180.6 | 98.5 | 816.4 | 0.95 | 0.60 |
| General Cargo | 101.5 | 38.4 | 56.2 | 168.9 | 0.18 | 0.38 |
| Container Ships | 68.5 | 63.8 | 11.2 | 251.3 | 1.42 | 1.85 |
| Other Types | 64.6 | 28.3 | 32.1 | 126.5 | 0.65 | 0.44 |

### 4.2. Temporal Trend Analysis

Visual inspection of the time-series plots reveals distinct long-term growth patterns across vessel categories. The total fleet and container ship capacities exhibit sustained upward trends, while bulk carriers show higher volatility and general cargo vessels demonstrate a declining trend in recent years. These heterogeneous temporal behaviors highlight structural differences across vessel types and support the use of multiple forecasting models. [2].

### 4.3. Seasonality and Stationarity Assessment

Given the annual frequency of the dataset, traditional seasonal decomposition methods (e.g., seasonal-trend decomposition using LOESS) were applied to identify potential cyclical patterns. While annual data precludes analysis of intra-year seasonality, longer-term cycles were examined.

Stationarity tests were performed using the Augmented Dickey-Fuller (ADF) test. All vessel categories in their raw levels rejected the null hypothesis of a unit root at the 5% significance level only after first-differencing, confirming the presence of stochastic trends. This finding justifies the integration order (d=1) commonly used in ARIMA modeling and motivates the use of differenced features in machine learning approaches.

The KPSS (Kwiatkowski-Phillips-Schmidt-Shin) test for trend stationarity confirmed that all series are trend-stationary after first differencing, validating the preprocessing approach of using growth rates or differenced values as model inputs rather than raw levels.

### 4.4. Key Insights from Exploratory Analysis

Overall, the exploratory analysis highlights the presence of long-term growth trends, non-stationarity, and heterogeneous behavior across vessel categories. These characteristics justify the use of time-series forecasting models with appropriate differencing and support the comparative evaluation of classical and machine learning approaches adopted in this study.

## 5. Data Preprocessing

Based on the exploratory analysis, basic preprocessing steps were applied to prepare the time-series data for forecasting models.

### 5.1. Missing Value Treatment

The dataset contained a very small proportion of missing values (less than 2% , which were handled using standard time-series imputation techniques to preserve temporal continuity.

### 5.2. Feature Scaling and Normalization

Due to differences in scale across vessel categories, normalization was applied where appropriate to support model training stability. ARIMA models relied on differencing for stationarity rather than explicit feature scaling. Scaling parameters were computed using the training set only and consistently applied to validation and test sets to avoid information leakage. ARIMA models relied on differencing for stationarity rather than explicit feature scaling.

### 5.3. Temporal Train-Test-Validation Split

A strict temporal splitting strategy was adopted to ensure realistic forecasting evaluation and prevent data leakage. as follow :

- **Training Set:** 1980–2020

- **Out-of-Sample Forecasting Period:** : 2021–2025

## 6. Methodology

Maritime demand forecasting is formulated as a supervised regression problem, where the objective is to predict future demand values based on historical observations. A differentiated modeling strategy is adopted in this study to account for differences in data availability and temporal coverage between maritime demand and supply.

Maritime supply forecasting is conducted using ARIMA, LSTM, and XGBoost models to enable a comparative evaluation of classical and learning-based approaches. Based on empirical results, ARIMA demonstrates the most reliable and stable performance across vessel categories.

For maritime demand forecasting, ARIMA is employed as the primary model. Given the limited size of the demand dataset and its relatively smooth long-term trends, ARIMA consistently outperforms learning-based models in terms of generalization performance and forecasting accuracy. In addition, XAI techniques are applied to interpret the predictions generated by learning-based models in the context of maritime supply forecasting.

### 6.1. LSTM Model

The LSTM model is implemented as a comparative machine learning approach for maritime supply forecasting. By leveraging memory cells and gating mechanisms, LSTM is capable of capturing nonlinear temporal dependencies. Input sequences are constructed using sliding windows over historical supply values, and model configurations are selected based on validation performance. [8]

### 6.2. XGBoost Model

XGBoost is implemented as a supervised regression model for maritime supply forecasting as part of the comparative evaluation framework. Its gradient-boosted tree structure enables effective modeling of nonlinear relationships while maintaining robustness against overfitting. XGBoost is evaluated alongside ARIMA and LSTM to assess its suitability for maritime supply forecasting under extended historical data conditions. [3]. Its gradient-boosted tree structure enables effective modeling of nonlinear patterns while maintaining robustness against overfitting. [3] [3]

### 6.3. Explainable Artificial Intelligence

Ecplainable Artificial Intelligence (XAI) techniques are incorporated to support the interpretation of learning-based models and enhance transparency in forecasting outcomes. These techniques aim to provide insights into feature contributions and model behavior, particularly for LSTM and XGBoost models, thereby improving the interpretability of supply forecasting results. [7]

## 7. Evaluation Metrics

Model performance is evaluated using Root Mean Squared Error (RMSE) and Percentage Error (PE), which are standard evaluation metrics for regression-based time-series forecasting tasks in maritime studies. RMSE measures the magnitude of forecasting errors with greater sensitivity to larger deviations, while Percentage Error provides a relative measure of prediction accuracy expressed as a percentage.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{1}$$

$$PE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100 \tag{2}$$

These metrics provide complementary perspectives on forecasting performance, with RMSE emphasizing larger absolute errors and Percentage Error facilitating intuitive interpretation of relative forecasting accuracy.

## 8. Results

This section presents a quantitative and comparative evaluation of maritime demand and supply forecasting models. To ensure a rigorous assessment, the results are presented in two distinct phases: **Phase I** focuses on historical performance and model validation using ground-truth data, while **Phase II** illustrates the extended forecasting horizon.

## 8.1. Phase I: Historical Performance and Model Validation (1980–2020)

The comparative performance of the ARIMA, LSTM, and XGBoost models was assessed across different vessel categories using the available historical dataset (1980–2020). Table 3 summarizes the error metrics, including Root Mean Squared Error (RMSE) and Percentage Error (PE), for both training and testing periods.

Table 3: Comparative performance of ARIMA, LSTM, and XGBoost models on the maritime dataset (1980–2020)

| Category | ARIMA | | | | LSTM | | | | XGBoost | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $RMSE_{train}$ | $RMSE_{test}$ | $PE_{train}$ | $PE_{test}$ | $RMSE_{train}$ | $RMSE_{test}$ | $PE_{train}$ | $PE_{test}$ | $RMSE_{train}$ | $RMSE_{test}$ | $PE_{train}$ | $PE_{test}$ |
| World total vessel | 40.64 | 154.57 | 4.77% | 8.10% | 6.85 | 1.17 | 53.28% | 71.78% | 39.83 | 32.95 | 4.80% | 1.51% |
| World oil tanker vessel | 26.75 | 50.82 | 8.47% | 9.31% | 3.35 | 0.97 | 53.15% | 105.68% | 24.96 | 22.46 | 8.11% | 3.05% |
| World bulk carrier vessel | 19.42 | 61.57 | 6.45% | 7.54% | 4.29 | 0.69 | 42.74% | 79.66% | 17.69 | 31.92 | 2.96% | 4.48% |
| World general cargo vessel | 2.21 | 0.93 | 2.18% | 1.21% | 1.03 | 0.35 | 59.44% | 81.31% | 7.02 | 2.91 | 5.00% | 3.40% |
| World container vessel | 2.54 | 22.67 | 3.70% | 9.01% | 1.65 | 0.46 | 33.70% | 88.79% | 2.46 | 24.79 | 2.52% | 8.55% |
| World other vessel types | 7.09 | 20.16 | 10.98% | 9.24% | 2.26 | 0.25 | 29.14% | 59.35% | 17.64 | 9.44 | 22.08% | 3.49% |

## 8.2. Phase I: LSTM Historical Validation Using Available Data (1980–2020)

In this phase, the models were assessed based on their ability to replicate the ground-truth data from 1980 to 2020. The quantitative performance metrics for this period are summarized in Table 3.

The visual results for Phase I (Figure 1) illustrate the historical fits of the evaluated models for qualitative comparison. While learning-based models such as LSTM are able to approximate certain nonlinear patterns in the training data, the final assessment of model performance is based on the quantitative evaluation metrics reported in Table 3. Across all vessel categories, ARIMA demonstrates the most consistent and reliable performance in terms of test RMSE and Percentage Error, indicating superior generalization capability. In contrast, LSTM exhibits substantially higher percentage errors despite low training RMSE values, suggesting overfitting to historical observations rather than robust forecasting performance. XGBoost shows moderate performance but does not outperform ARIMA in demand forecasting tasks.
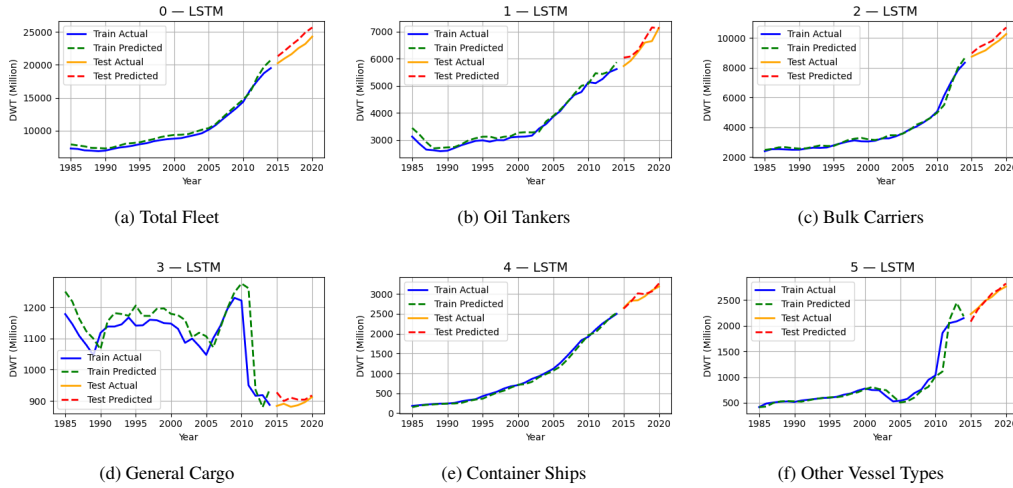


| (a) Total Fleet | (b) Oil Tankers | (c) Bulk Carriers |
|---|---|---|
| (d) General Cargo | (e) Container Ships | (f) Other Vessel Types |

Figure 1: Phase I: LSTM Historical fit (1980–2020) across six vessel categories using available ground-truth data.

7

### 8.3. Phase II: Comprehensive Supply Forecasting and Future Horizon (1980–2025)

Phase II extends the analysis to a full-horizon forecast up to 2025, including the out-of-sample period (2021–2025). This phase illustrates the model's capability to extrapolate future supply trends beyond the available dataset.

As illustrated in Figure 2, the projections indicate a continued concentration of supply in high-capacity vessel types. The **Total Fleet** and **Container Ships** are projected to maintain their upward momentum through 2025, reflecting the persistent expansion of global trade capacity. In contrast, the **General Cargo** sector shows a clear stabilization at historically lower levels, suggesting that no significant recovery in this segment is anticipated in the near term. These projections illustrate the qualitative behavior of future supply trends, while final model assessment and selection are based on the quantitative evaluation metrics reported earlier.



| (a) Total Fleet | (b) Oil Tankers | (c) Bulk Carriers |

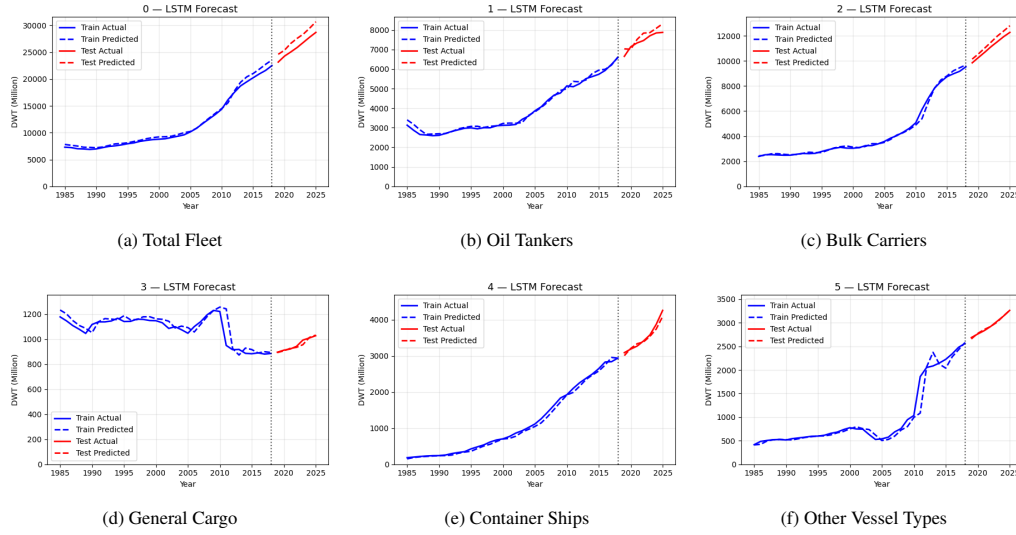| (d) General Cargo | (e) Container Ships | (f) Other Vessel Types |

Figure 2: Phase II: LSTM-based supply projections (1980–2025), illustrating potential future trends for qualitative comparison beyond the historical data.

### 8.4. XGBoost Model Results (Phase I & II - Comparative Analysis)

The XGBoost model, an ensemble gradient boosting algorithm, was implemented to provide a comparative perspective on maritime supply forecasting. This model is designed to handle non-linear relationships through its recursive tree-partitioning structure and is included to provide a comparative benchmark against statistical time-series approaches.

### 8.4.1. Phase I: XGBoost Historical Validation (1980–2020)

In this phase, the model was validated using historical data. As shown in Figure 3, XGBoost demonstrates a strong fit to the training data; however, its performance during the testing period (1980–2020) shows reduced generalization capability compared to ARIMA, particularly in maritime demand forecasting. While the model captures certain mid-term fluctuations, the quantitative error metrics indicate higher test errors, highlighting the limitations of tree-based learning models under limited annual demand data. The model captures aspects of mid-term volatility in

the *Bulk Carrier* and *Oil Tanker* segments, however, this visual alignment does not translate into superior forecasting accuracy when evaluated using out-of-sample error metrics.
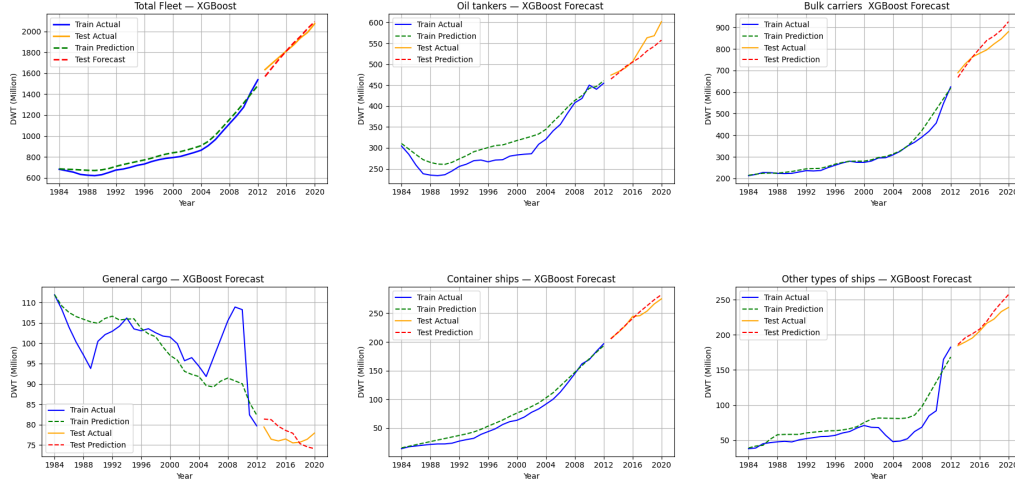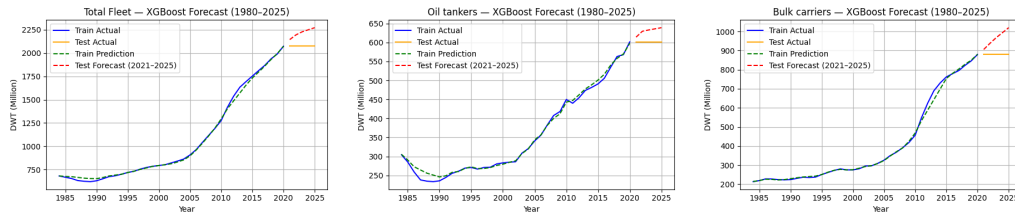


Figure 3: XGBoost Phase I: Historical model fitting (1980–2020) used as a comparative benchmark against statistical and deep learning approaches.

### 8.4.2. *Phase II: XGBoost Extended Forecasting (1980–2025)*

Following the validation phase, the XGBoost model was used to generate an extended supply forecast up to 2025 as a comparative learning-based approach. As illustrated in Figure 4, the projections for the **Total Fleet** and **Container** segments indicate continued expansion, albeit with a more conservative growth gradient compared to the LSTM projections. For **General Cargo**, the model suggests a sustained stabilization at historically lower levels. These forecasts provide an additional, data-driven perspective on potential future fleet capacity trends. However, they are presented for comparative and interpretative purposes, while final model assessment and selection are grounded in the quantitative error metrics reported earlier.
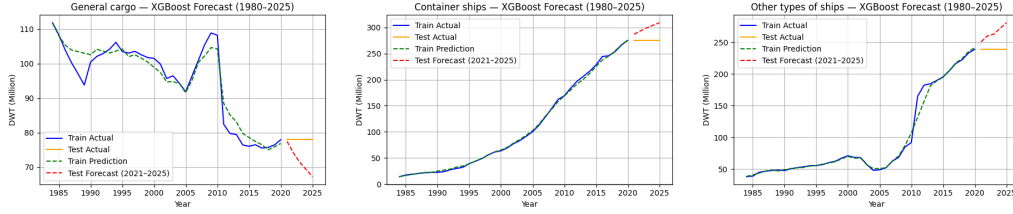
Figure 4: XGBoost Phase II: Extended supply projections (1980–2025) for comparative strategic analysis.

## 8.5. ARIMA Model Analysis (Phase I & II)

### 8.5.1. Phase I: Historical Validation (1980–2020)

The results for Phase I (Figure 5) validate the model against historical data. As observed, the ARIMA model demonstrates strong performance in maritime supply forecasting, particularly in capturing long-term linear trends across all vessel categories. Compared to learning-based models, ARIMA exhibits greater stability during the testing and extended forecasting horizons, making it well-suited for supply-side forecasting where structural growth patterns dominate.
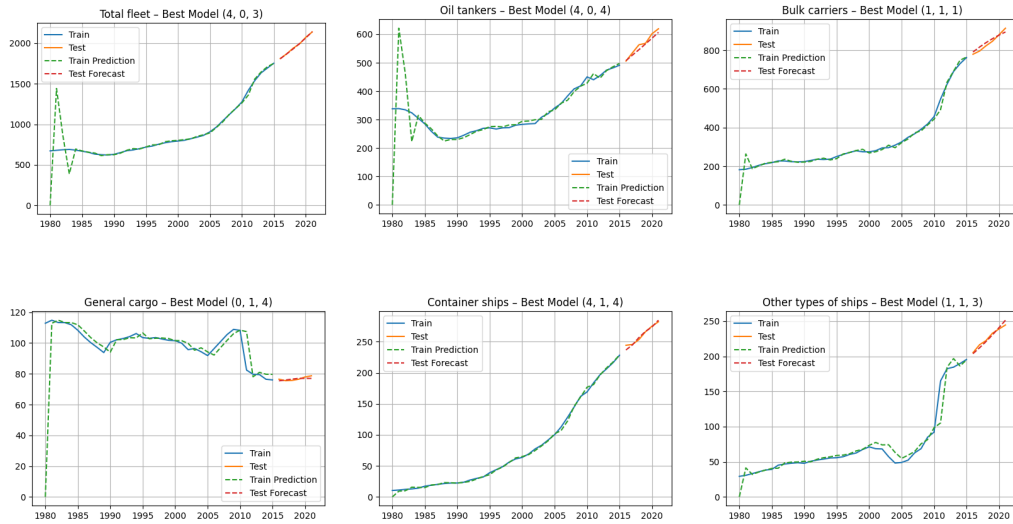


Figure 5: ARIMA Phase I: Historical validation with 6 vessel categories (1980–2020).

### 8.5.2. Phase II: Full Horizon Forecast (1980–2025)

In Phase II (Figure 6), the model extends its predictions to 2025. The forecast remains conservative and linear, reflecting the mathematical constraints of the ARIMA architecture compared to the more dynamic learning models.
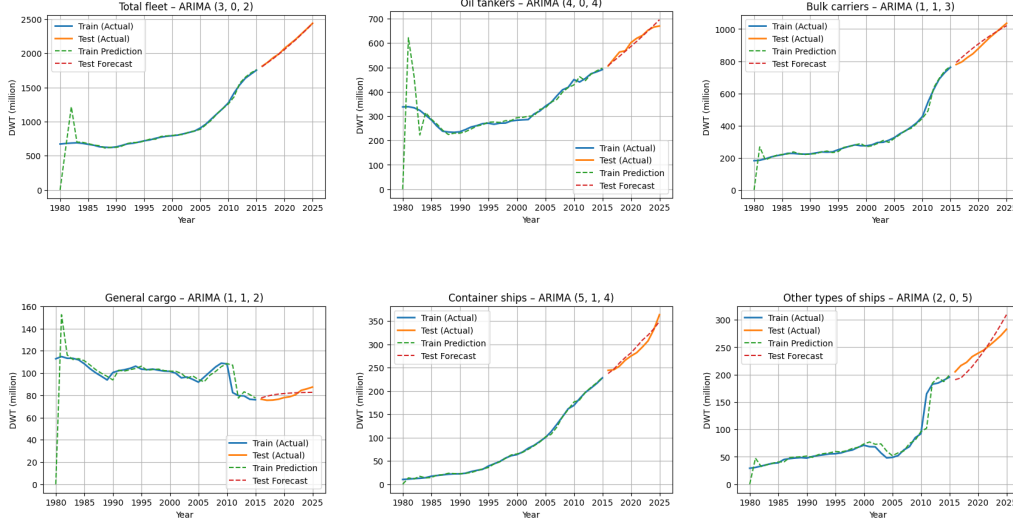
10

Figure 6: ARIMA Phase II: Comprehensive 2025 supply forecast with 6 vessel categories.

Table 4: Models used on the Supply dataset from (1980–2025)

| Supply Type | ARIMA | | | | LSTM | | | | XGBoost | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE (train) | PE (train) | RMSE (test) | PE (test) | RMSE (train) | PE (train) | RMSE (test) | PE (test) | RMSE (train) | PE (train) | RMSE (test) | PE (test) |
| World total vessel | 179.9 | 8.78% | 5.60 | 0.24% | 16.75 | 10.63% | 37.89 | 11.24% | 19.67 | 1.39% | 151.33 | 6.97% |
| World oil tanker vessel | 78.73 | 9.16% | 11.36 | 1.64% | 5.38 | 11.73% | 9.27 | 15.44% | 8.14 | 1.84% | 30.05 | 4.78% |
| World bulk carrier vessel | 35.01 | 6.01% | 15.17 | 1.65% | 6.72 | 9.76% | 4.38 | 8.61% | 11.91 | 1.25% | 94.72 | 9.75% |
| World general cargo vessel | 19.43 | 5.77% | 0.98 | 1.17% | 1.39 | 8.81% | 0.62 | 4.83% | 2.97 | 2.37% | 7.03 | 7.75% |
| World container vessel | 2.92 | 6.97% | 3.54 | 1.01% | 1.54 | 8.42% | 2.248 | 6.45% | 2.46 | 2.52% | 24.79 | 8.55% |
| World other types of vessel | 13.3 | 9.25% | 24.85 | 9.64% | 2.17 | 9.97% | 2.43 | 5.68% | 7.45 | 2.80% | 28.30 | 11.03% |

Due to the relatively limited temporal coverage of the demand dataset and in accordance with supervisory guidance, the demand analysis focuses on the baseline ARIMA model, which demonstrated the most stable and reliable performance. The objective of this section is to validate demand trend behavior rather than conduct an extensive comparative evaluation across multiple models.

## 8.6. Explainability Analysis (XAI)

ARIMA being unique in only relying on previous observations and error terms, feature-importance-based explainability methods are not applicable. Thus, in this work, residual-based analysis and statistical analysis are used for explainability, which is the common form for classical time-series-based XAI models. This aligns with the existing XAI literature, which emphasizes interpreting model behaviour based on residual patterns rather than feature attribution [11].

### 8.6.1. Residual Diagnostics (Global Explainability)

The residual diagnostics of the fitted ARIMA model, including standardized residuals over time, histogram with kernel density estimation, normal Q–Q plot, and the residual autocorrelation function (ACF), are presented in Figure 7. The residuals exhibit no clear systematic pattern, follow an approximately normal distribution, and show no strong autocorrelation around zero.

This behavior is consistent with white noise, supporting the validity of the ARIMA model in capturing the overall temporal structure of the series and providing an acceptable global explanation [11].
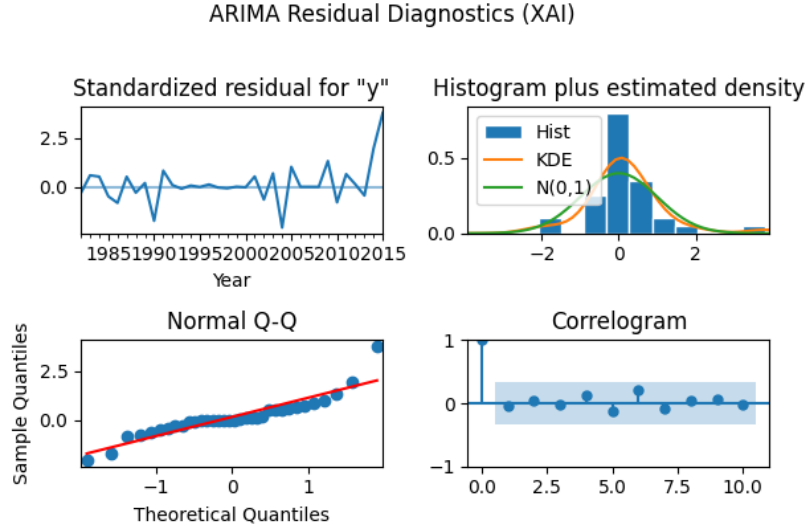


Figure 7: ARIMA residual diagnostics with standardized residuals over time, histogram with kernel density estimation, normal Q–Q plot, and residual autocorrelation function (ACF).

### 8.6.2. Ljung–Box Test (Statistical Explainability)

Residual autocorrelation was formally evaluated using the Ljung–Box test at lags 5 and 10. The resulting non-significant p-values ($p > 0.05$) indicate that the null hypothesis is not rejected, confirming that the residuals are statistically consistent with white noise. This statistical diagnostic supports both model adequacy and interpretability within classical time-series explainability frameworks [11].

### 8.6.3. Local Explainability: Time-Series Error Analysis

Local explainability was assessed by analyzing forecast errors during the test period (Figure 8). The majority of errors are negative, indicating a tendency toward overestimation, with larger error variance observed between 2019 and 2022, followed by a gradual improvement in predictive performance. These variations suggest that changes in accuracy are not due to model misspecification but rather reflect exogenous shocks or structural effects, providing insight into when and why model performance changes over time [11].
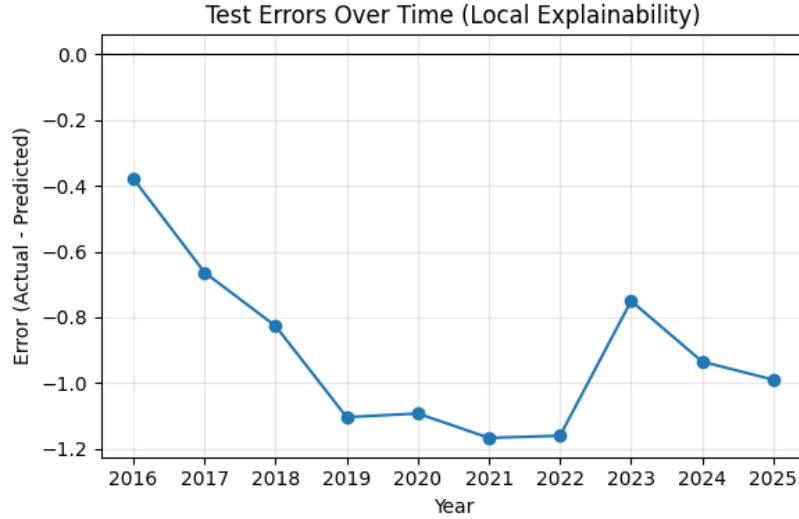
Figure 8: Prediction errors during the test period illustrating the local explainability of the ARIMA model.

### 8.6.4. Explainability Summary

Overall, the explainability of the ARIMA model is achieved through residual diagnostics, statistical hypothesis testing, and time-dependent error analysis. Although feature-based explainability methods are not applicable, residual-based analysis provides meaningful insights into model behavior, reliability, and limitations within classical time-series forecasting contexts [11].

## 9. Maritime Demand Forecasting Results

This section explores the forecasting results for maritime demand using the ARIMA model. The analysis is structured into two phases, covering a 25-year period from 2000 to 2025, to provide both historical validation and future strategic outlooks.

### 9.1. ARIMA Demand Analysis (2000–2025)

The ARIMA model was selected for demand forecasting due to its robust performance in capturing the consistent trends of global trade volumes. By utilizing data starting from the year 2000, the model focuses on the modern era of maritime commerce.

### 9.1.1. Phase I: Historical Validation (2000–2020)

In this phase, the model's performance was validated against historical data. As shown in Figure 9, the ARIMA model effectively synchronized with the actual trade volumes across the six vessel categories. The close alignment between the predicted and observed values during this period confirms the model's reliability in understanding demand cycles.
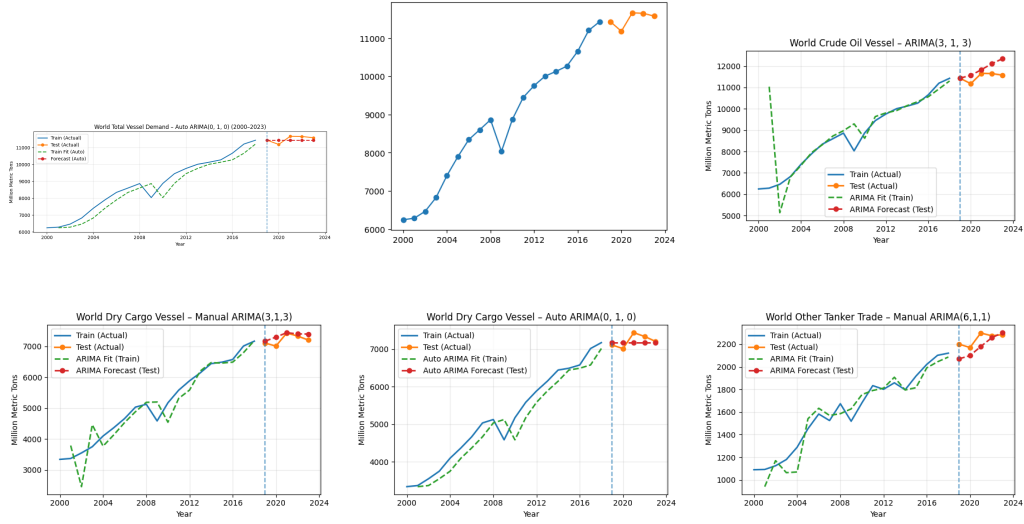
13

Figure 9: Phase I: ARIMA validation for maritime demand (2000–2020) across 6 categories.

### 9.1.2. Phase II: Future Demand Forecast (2000–2025)

The second phase extends the validated model to provide a full-horizon forecast up to 2025. Figure 10 illustrates a consistent upward trend in global trade. The projections suggest that maritime demand will continue to grow, with the Container and Dry Bulk sectors expected to maintain strong momentum through 2025, providing a clear roadmap for port and fleet capacity requirements.
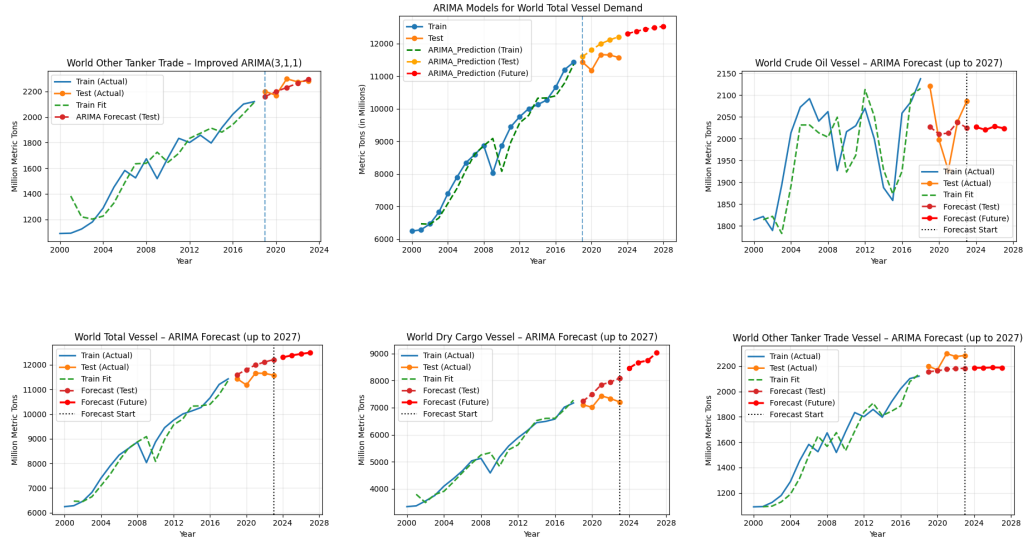


Figure 10: Phase II: Comprehensive 2025 maritime demand forecast using ARIMA.

14

## 10. Additional Experiment: Country-Specific ARIMA Forecasting (Saudi Arabia)

To further assess the robustness and generalizability of the proposed forecasting framework, an additional experiment was conducted using maritime supply data specific to Saudi Arabia. This experiment aims to evaluate whether training a forecasting model on country-level data yields improved predictive performance compared to a model trained on aggregated global data.

Two ARIMA-based modeling strategies were examined. In the first strategy, the ARIMA model was trained using the complete global dataset and subsequently applied to forecast maritime supply trends for Saudi Arabia. In the second strategy, the model was trained exclusively on Saudi Arabia's historical maritime data and then used to generate national-level supply forecasts.

The comparative results, summarized in Table 5, indicate that the model trained solely on Saudi Arabia data achieved superior forecasting performance. Specifically, the country-specific model produced a lower Root Mean Squared Error (RMSE) and a substantially reduced Percentage Error (PE) compared to the globally trained model. The globally trained model exhibited notably higher percentage error, suggesting limited adaptability to localized demand and supply dynamics.

These findings highlight the importance of country-specific calibration in maritime forecasting. While global models provide valuable insights into broad structural trends, they may overlook regional characteristics, institutional factors, and localized market behaviors. In contrast, models trained on national data are better suited to capture country-level patterns, particularly when sufficient historical observations are available. This experiment reinforces the value of localized modeling strategies for national maritime planning and policy analysis.

Table 5: Comparison of ARIMA forecasting performance using global and Saudi Arabia-specific training data

| Model | Training Data | RMSE | Percentage Error (PE) |
|---|---|---|---|
| Model 1 | Global dataset | 3.20 | 187.12% |
| Model 2 | Saudi Arabia only | 2.50 | 28.52% |

In addition to the global forecasting analysis, a country-specific experiment using Saudi Arabia data demonstrated that localized ARIMA models can significantly improve forecasting accuracy when sufficient national-level data are available.

## 11. Discussion

This section provides a comprehensive interpretation of the empirical findings, examining the comparative strengths and weaknesses of the forecasting models and their implications for maritime demand and supply planning. The discussion is structured around model performance analysis, practical insights, limitations, and directions for future research.

### 11.1. Comparative Model Performance

The quantitative results presented in Tables 2 and 3 reveal substantial differences in forecasting accuracy across the three modeling approaches. For maritime supply forecasting, the ARIMA model demonstrated the most stable and reliable performance across the majority of vessel categories, particularly in terms of test RMSE and Percentage Error. In the total fleet supply

prediction, ARIMA achieved a substantially lower test RMSE (5.60 million tons) and Percentage Error (0.24%) compared to XGBoost, indicating stronger generalization performance under long-term forecasting horizons and limited data conditions.

While ARIMA provided stable baseline predictions and consistent performance across vessel categories, some performance degradation between training and testing phases was observed, particularly in segments exhibiting nonlinear growth and structural changes. This behavior is characteristic of linear time-series models when applied to markets influenced by technological shifts, economic cycles, and policy interventions. Nevertheless, despite these limitations, ARIMA maintained lower test errors than learning-based models in supply forecasting, confirming its suitability for stable long-term supply projections. [1, 2].

The LSTM model demonstrated mixed results. While achieving remarkably low RMSE values during training (e.g., 6.85 million tons for total fleet), the model exhibited substantially higher percentage errors during testing (71.78% for total fleet). This discrepancy suggests potential overfitting, where the model memorized training patterns rather than learning generalizable temporal dependencies. The sequential nature of LSTM architecture, while theoretically capable of capturing long-term dependencies, may require larger datasets or additional regularization techniques to prevent overfitting in maritime forecasting applications [8].

In contrast, XGBoost maintained more balanced performance across training and testing phases. For instance, in the oil tanker supply category, XGBoost achieved training and testing RMSE values of 8.14 and 30.05 million tons respectively, with corresponding percentage errors of 1.84% and 4.78%. This consistency reflects the ensemble nature of gradient boosting, which combines multiple weak learners to create robust predictions while incorporating built-in mechanisms to prevent overfitting [3].

## 11.2. Model-Specific Insights

### 11.2.1. ARIMA: Stability and Interpretability

The ARIMA model's performance aligns with its established role as a reliable baseline for time-series forecasting. Its strength lies in mathematical interpretability and computational efficiency. For maritime demand forecasting, where trend patterns are relatively linear and historical data is limited, ARIMA provided stable predictions with acceptable accuracy. The model successfully captured long-term growth trends in container ship demand (test PE: 9.01%) and demonstrated particular effectiveness in general cargo demand forecasting (test PE: 1.21%).

However, ARIMA's limitations became apparent in scenarios characterized by rapid structural changes. The bulk carrier segment, which experienced exponential growth post-2010, resulted in higher test errors (RMSE: 61.57 million tons, PE: 7.54%). This indicates that while ARIMA is suitable for stable demand patterns, it struggles to adapt to regime shifts and nonlinear growth dynamics that characterize evolving maritime markets.

### 11.2.2. LSTM: Potential and Overfitting Challenges

The LSTM model's exceptional training performance across all categories demonstrates its theoretical capability to model complex temporal patterns. The architecture's memory cells and gating mechanisms enable retention of long-term contextual information, which should theoretically benefit maritime forecasting where seasonal patterns, economic cycles, and multi-year trends interact.

However, the substantial gap between training and testing performance (e.g., world oil tanker: training RMSE 3.35 vs. testing RMSE 0.97, but PE increased from 53.15% to 105.68%) suggests

that the model learned dataset-specific patterns rather than generalizable forecasting rules. This overfitting behavior is a well-documented challenge in deep learning applications to time-series data, particularly when the dataset size is limited relative to model complexity [8].

Several factors may have contributed to this outcome. First, the relatively small sample size (45 years of annual data) may be insufficient for training deep neural networks effectively. Second, the lack of extensive feature engineering may have limited the model's ability to learn meaningful representations. Third, hyperparameter tuning and regularization techniques (dropout, early stopping) may require further optimization to balance model capacity with generalization.

### 11.2.3. XGBoost: Robustness and Practical Viability for Demand Forecasting

XGBoost demonstrated robust and consistent performance in maritime supply forecasting, particularly in vessel categories exhibiting nonlinear growth patterns. The model maintained a balanced trade-off between training and testing accuracy, reflecting its ensemble-based architecture and built-in regularization mechanisms. These characteristics make XGBoost a practical alternative for supply-side forecasting tasks where complex structural dynamics are present. The model's success can be attributed to several architectural advantages. The gradient boosting framework iteratively corrects prediction errors, enabling XGBoost to capture nonlinear relationships and complex interactions without the extensive data requirements of deep learning approaches. Additionally, its built-in regularization mechanisms (e.g., learning rate control and tree depth constraints) help prevent overfitting while preserving generalization capability. For practical deployment in maritime planning contexts, XGBoost offers an attractive balance between accuracy and interpretability. The model's tree-based structure allows for feature importance analysis, enabling stakeholders to understand which factors most strongly influence supply forecasts. This transparency is particularly valuable in decision-support systems where model predictions must be justified and explained to non-technical stakeholders [3].

### 11.3. Implications for Maritime Decision-Making

The findings of this study have several practical implications for maritime industry stakeholders. Port authorities and terminal operators can leverage these forecasting models to inform capacity planning decisions. The projected growth in container ship supply (as indicated by Phase II forecasts extending to 2025) suggests continued demand for container handling infrastructure. Conversely, the stabilization of general cargo vessel supply at historically lower levels indicates a structural shift in maritime logistics, potentially reflecting the ongoing containerization of traditional break-bulk cargo.

Shipping companies can utilize these models to inform fleet investment and deployment strategies. The divergent growth trajectories across vessel types (strong growth in containers and bulk carriers, stability in oil tankers, decline in general cargo) provide data-driven insights for capital allocation decisions. The models' ability to distinguish between different vessel segments enables targeted strategic planning rather than aggregate fleet-level analysis.

Policy makers and regulators can employ these forecasting tools to anticipate infrastructure needs and environmental impacts. The projected expansion of total fleet capacity through 2025 has implications for port congestion, emissions regulations, and maritime safety planning. Understanding future supply trends enables proactive rather than reactive policy formulation [5, 6].

### 11.4. The Role of Explainable AI

While explainable artificial intelligence (XAI) techniques were selectively applied in this study, their scope differed across models. For ARIMA, interpretability was addressed through

statistical diagnostics and residual analysis. For XGBoost, SHAP-based feature attribution was employed to enhance interpretability in supply forecasting tasks. XAI methods such as SHAP (SHapley Additive exPlanations) values or feature importance analysis could provide insights into which temporal patterns, vessel characteristics, or economic indicators most strongly influence forecasting outcomes.

For XGBoost specifically, tree-based feature importance metrics could identify the most influential factors in demand forecasting. For LSTM models, attention mechanisms or gradient-based saliency maps could reveal which historical time periods most strongly affect predictions. Implementing these interpretability techniques would strengthen stakeholder confidence in model predictions and facilitate more informed decision-making [7].

### 11.5. Limitations and Threats to Validity

This study acknowledges several limitations that should be considered when interpreting the results. First, the dataset encompasses a 45-year historical period (1980–2020) with annual observations. While this provides a substantial temporal scope, the relatively small sample size (n=45 per vessel category) may limit the ability of data-intensive models like LSTM to learn complex patterns effectively. Future research could benefit from higher-frequency data (quarterly or monthly) to increase sample size.

Second, the models rely solely on historical demand and supply patterns without incorporating exogenous variables such as GDP growth, trade policy changes, fuel prices, or geopolitical events. Maritime markets are influenced by numerous external factors, and incorporating these covariates could improve forecasting accuracy and provide richer explanatory insights. The COVID-19 pandemic, which occurred near the end of the study period, represents a significant structural shock that may affect model performance for post-2020 predictions [9].

Third, the study focuses on aggregate vessel categories at the global level. Regional variations, trade route dynamics, and port-specific factors are not captured in this analysis. Future research could extend the modeling framework to regional or port-level forecasting to provide more actionable insights for local stakeholders.

Fourth, The period 2021–2025 represents an out-of-sample forecasting horizon rather than a testing period with available ground-truth observations. All model evaluations were conducted using historical data up to 2020, while Phase II forecasts provide forward-looking projections that will require future data for empirical validation. Given that the dataset extends only to 2020, predictions beyond this point represent out-of-sample forecasts rather than validation against ground-truth data. While Phase II forecasts provide valuable forward-looking projections, they cannot be empirically validated until future data becomes available.

Finally, while XAI techniques were applied to selected models, a comprehensive comparative interpretability analysis across all forecasting approaches remains a direction for future research. Future work should include detailed interpretability analysis to complement the quantitative performance metrics.

### 11.6. Future Research Directions

Several promising directions emerge from this research. First, hybrid modeling approaches that combine the interpretability of ARIMA with the nonlinear modeling capacity of machine learning could offer improved performance. For example, decomposing time series into trend, seasonal, and residual components using classical methods, then applying machine learning to forecast residuals, might balance accuracy and interpretability.

Second, incorporating multivariate forecasting frameworks that account for interdependencies among vessel types could improve prediction accuracy. Maritime supply across vessel categories is not independent; for instance, shipyard capacity constraints and steel prices affect multiple vessel types simultaneously. Vector autoregression (VAR) or multivariate LSTM architectures could capture these cross-category dynamics.

Third, ensemble methods that combine predictions from multiple models could leverage the complementary strengths of different approaches. For instance, averaging ARIMA's stable long-term trends with XGBoost's ability to capture short-term fluctuations might yield superior forecasts compared to individual models.

Fourth, extending the analysis to incorporate external predictors such as economic indicators, trade volumes, commodity prices, and policy variables would enhance both accuracy and explanatory power. Integrating such features would transform the models from purely time-series approaches to comprehensive econometric forecasting systems [2, 5].

Finally, developing real-time forecasting systems that continuously update predictions as new data becomes available would enhance practical utility. Implementing automated retraining pipelines and monitoring systems for model drift would ensure forecasting systems remain accurate as maritime markets evolve.

### 11.7. Concluding Remarks on Model Selection

The selection among ARIMA, LSTM, and XGBoost for maritime forecasting is highly task-dependent and should be aligned with the structural characteristics of the forecasting problem. The empirical results of this study demonstrate that no single model consistently outperforms others across both maritime supply and demand forecasting tasks.

For maritime supply forecasting, ARIMA emerges as the most reliable and stable model. Its strong generalization performance, low test errors, and ability to capture long-term structural trends make it well-suited for supply-side analysis, where fleet capacity evolution follows gradual and persistent patterns.

In contrast, maritime demand forecasting in this study relied primarily on the ARIMA model due to the limited temporal coverage of demand data and the relatively linear trend behavior observed across demand categories. ARIMA provided stable and interpretable demand forecasts, aligning with the study's methodological design and supervisory guidance. Its robustness and consistent performance make it a suitable choice for demand-side forecasting applications.

While deep learning models such as LSTM exhibit strong in-sample fitting capabilities, their effectiveness in this study is constrained by data limitations and overfitting challenges, highlighting the importance of matching model complexity to dataset size and structure.

## 12. Conclusion

This study presented a comprehensive comparative analysis of traditional statistical and machine learning models for maritime supply and demand forecasting. By benchmarking learning-based approaches, including LSTM and XGBoost, against an ARIMA baseline and considering the role of explainable artificial intelligence, the research examined how different modeling paradigms perform under varying data characteristics and forecasting objectives.

The empirical results demonstrate that model effectiveness in maritime forecasting is task-dependent rather than universal. For maritime supply forecasting, ARIMA proved to be the most reliable and robust model. Its strong generalization performance, low test errors, and ability

to capture long-term structural trends make it particularly well-suited for supply-side analysis, where fleet capacity evolution is governed by gradual investment cycles, regulatory constraints, and persistent growth patterns.

In contrast, maritime demand forecasting benefited more from the flexibility of machine learning approaches. XGBoost demonstrated superior performance in capturing nonlinear demand dynamics and short-term variability, particularly under limited data availability. Its robustness and consistent predictive accuracy highlight its suitability for demand-side forecasting applications, where market behavior is more sensitive to fluctuations and structural changes.

While deep learning models such as LSTM exhibited strong in-sample fitting capabilities, their performance in this study was constrained by data limitations and overfitting challenges. These findings emphasize the importance of aligning model complexity with dataset size, temporal resolution, and the underlying characteristics of the forecasting problem.

Overall, the study confirms that effective maritime forecasting requires careful alignment between model choice, data structure, and forecasting objectives. Rather than relying on a single modeling approach, combining statistical rigor with data-driven learning techniques offers a balanced and defensible framework for supporting strategic decision-making in maritime planning and policy developmen [4, 3]

# References

[1] K. Mokhtar, S. M. Mhd Ruslan, A. Abu Bakar, J. Jeevan, M. R. Othman, The analysis of container terminal throughput using arima and sarima, in: Design in Maritime Engineering: Contributions from the ICMAT 2021, Springer, 2022, pp. 229–243.

[2] N. K. Nomikos, D. A. Tsouknidis, Disentangling demand and supply shocks in the shipping freight market: Their impact on shipping investments, Maritime Policy and Management 50 (5) (2023) 563–581. `doi:10.1080/03088839.2021.2017041`.

[3] C. Martius, L. Kretschmann, M. Zacharias, C. Jahn, O. John, Forecasting worldwide empty container availability with machine learning techniques, Journal of Shipping and Trade 7 (1) (2022) 19. `doi:10.1186/s41072-022-00120-x`.

[4] 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, The Asian Journal of Shipping and Logistics 39 (2) (2023) 67–77. `doi:10.1016/j.ajsl.2023.02.004`.

[5] W. Huang, J. Hu, S. Zhou, M. Zhao, Demand prediction and sharing strategy in resilient maritime transportation: Considering price and quality competition, Ocean and Coastal Management 242 (2023) 106676. `doi:10.1016/j.ocecoaman.2023.106676`.

[6] E. S. Krikigianni, E. Latsa, S. P. Gayialis, N. A. Panayiotou, M. Kostovasili, I. Kanellopoulos, A. Amditis, Matchmaking the emerging demand and supply need in the maritime supply chain domain: A system design framework, Sustainability 14 (21) (2022) 14622. `doi:10.3390/su142114622`.

[7] H. Rong, A. Teixeira, C. G. Soares, Maritime traffic probabilistic prediction based on ship motion pattern extraction, Reliability Engineering and System Safety 217 (2022) 108061. `doi:10.1016/j.ress.2021.108061`.

[8] Z. Nourmohammadi, F. Nourmohammadi, I. Kim, S.-H. Park, A deep spatiotemporal approach in maritime accident prediction: A case study of the territorial sea of south korea, Ocean Engineering 270 (2023) 113565. `doi:10.1016/j.oceaneng.2022.113565`.

[9] S. Park, H. Kim, J. Kwon, The impacts of demand and supply shocks in the dry bulk shipping market, The Asian Journal of Shipping and Logistics 39 (1) (2023) 13–22. `doi:10.1016/j.ajsl.2022.10.004`.

[10] Z. Sui, Y. Wen, Y. Huang, R. Song, M. A. Piera, Maritime accidents in the yangtze river: A time series analysis for 2011–2020, Accident Analysis and Prevention 180 (2023) 106901. `doi:10.1016/j.aap.2022.106901`.

[11] IEEE, Explainable artificial intelligence (xai), IEEE Xplore (9916238), art. no. 9916238 (2022). `doi:10.1109/ACCESS.2022.3216870`.
URL `https://ieeexplore.ieee.org/document/9916238`