City Research Online

City, University of London Institutional Repository

**Citation**: Alizadeh-Masoodian, A., Groven, B. R, Marchese, M., Moutzouris, I., Risstad, M. & Rustad, C. A. B. (2025). A hybrid combination approach to forecast freight rates volatility. Quantitative Finance, pp. 1-22. doi: 10.1080/14697688.2025.2568045

This is the published version of the paper.

This version of the publication may differ from the final published version.

**Permanent repository link:** https://openaccess.city.ac.uk/id/eprint/36099/

**Link to published version**: https://doi.org/10.1080/14697688.2025.2568045

**Copyright:** City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

**Reuse:** Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Quantitative Finance

ISSN: 1469-7688 (Print) 1469-7696 (Online) Journal homepage: www.tandfonline.com/journals/rquf20

# A hybrid combination approach to forecast freight rates volatility

A. Alizadeh, B. R. Groven, M. Marchese, I. Moutzouris, M. Risstad & C. A. B. Rustad

**To cite this article:** A. Alizadeh, B. R. Groven, M. Marchese, I. Moutzouris, M. Risstad & C. A. B. Rustad (22 Oct 2025): A hybrid combination approach to forecast freight rates volatility, Quantitative Finance, DOI: 10.1080/14697688.2025.2568045

**To link to this article:** https://doi.org/10.1080/14697688.2025.2568045

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 22 Oct 2025.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=rquf20

*Quantitative Finance*, 2025
https://doi.org/10.1080/14697688.2025.2568045

Routledge
Taylor & Francis Group
Check for updates

# A hybrid combination approach to forecast freight rates volatility

A. ALIZADEH†, B. R. GROVEN‡, M. MARCHESE†, I. MOUTZOURIS†, M. RISSTAD*‡ and C. A. B. RUSTAD‡

†Bayes Business School (formerly Cass), City St Georges, University of London, London, UK
‡Norwegian University of Science and Technology, Institute of Industrial Economics and Technology Management, Trondheim, Norway

(*Received 1 November 2024; accepted 24 September 2025*)

The aim of this paper is to investigate the performance of machine learning algorithms along with traditional GARCH and GARCH-MIDAS models in forecasting volatility of dry bulk shipping freight rates, known as one of the most volatile asset classes. In doing so, we introduce a new market tightness index, capturing physical constraints in shipping markets as an explanatory variable. The results suggest that significant incremental information can be extracted by Machine Learning algorithms from additional volatility predictors with minimal noise fitting, if regularization is applied. However, traditional GARCH models perform better in capturing the long-term persistence of the volatility. Therefore, a novel hybrid ensemble stacking algorithm that combines GARCH models and tree-based algorithms is proposed. This hybrid model, which utilizes exogenous predictors and the GARCH-MIDAS specification with the marked tightness index, produces accurate and robust out-of-sample volatility forecasts over a range of time horizons, from one day to one month.

*Keywords*: Volatility forecasting: Machine learning; CatBoost; Random forest; GARCH-MIDAS; Forecast combination: Freight rate; Shipping

## 1. Introduction

International shipping is the primary means of global transportation connecting production and consumption areas for raw materials and manufactured goods around the world, contributing to about 80% of the volume of international trade (UNCTAD 2023).† Shipping is also a complex industry as it encompasses four interconnected markets, namely the freight, sale and purchase, shipbuilding, and demolition markets. These interrelated markets define the supply and demand characteristics of the shipping freight market and create a highly cyclical and unpredictable environment where participants aim to balance risk and reward. The high sensitivity of shipping freight market to seasonal trade in commodities, port and canal congestion, macroeconomic shocks, oil and fuel prices, political events, weather and climate conditions and other factors moving the supply-demand equilibrium, results in a distinct and highly volatile freight market.

Given the high level of volatility in shipping freight markets and the importance of understanding their nature and behavior to market participants, several studies are devoted to modeling volatility of freight rates and asset prices. Almost all of these studies use different extensions of GARCH models to investigate dynamics of freight rate volatility. Kavussanos (1997) assess the efficiency of freight futures and forward freight agreements in hedging dry bulk freight rate risk (Kavussanos and Nomikos 2000, Kavussanos and Visvikis 2004), and explore the effect of market conditions and macroeconomic factors on freight rate volatility (Alizadeh and Nomikos 2011, Xu *et al.* 2011, Drobetz *et al.* 2012). Despite the importance of volatility forecasts and their use in freight risk assessment and management, there has been limited work on forecasting freight market volatility and evaluating volatility forecasts. The only exceptions are Abouarghoub *et al.* (2014), Gavriilidis *et al.* (2018), Argyropoulos and Panopoulou (2018), and Liu *et al.* (2022), who

*Corresponding author. Email: morten.risstad@ntnu.no

† The importance of international shipping has increased in recent years due to stronger economic ties between nations and significant growth in international trade. In addition, outsourcing production and manufacturing as well as discovery and production of raw materials in different parts of the world have contributed to the increase in ocean transportation. According to Clarksons' Shipping Intelligence Network (2024), it is estimated that, in 2023, 12 332 million tonnes of cargo were transported by various ship types, including dry bulk carriers, tankers, container ships, and gas carriers.

© 2025 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

utilize different forms of GARCH models to forecast volatility of shipping freight rates in tanker and dry bulk markets and estimate the corresponding Value-at-Risk. However, they fail to take into account the short- and long-term components of volatility, and their volatility prediction relies on limited or non-existent industry variables.

Many different types of GARCH models have also been proposed and utilized for modeling and forecasting economic and financial time series in other markets, including stock prices (e.g. Franses and Van Dijk 1996, Hansen and Lunde 2005, Anderson *et al.* 2006), foreign exchange (Boudt *et al.* 2013), and commodity prices (e.g. Agnolucci 2009, Y. Wang and Wu 2012, Bentes 2015, Herrera *et al.* 2018), to mention a few. However, recent studies on modeling and forecasting volatility recognize the effect of economic factors as well as short- and long-term components of volatility rather than relying on simple autoregressive specifications. For instance, Engle *et al.* (2013) propose a GARCH model with mixed data sampling (GARCH-MIDAS) to allow for the effect of macroeconomic variables with different frequencies (e.g. inflation and industrial production) on short- and long-run volatility of stock market. They report that the GARCH-MIDAS model incorporating macroeconomic factors can outperform simple GARCH models for short-term forecast of stock market volatility. L. Wang *et al.* (2020) extend the GARCH-MIDAS framework to allow for asymmetric and significant volatility effects due to extreme shocks to short- and long-term volatility components. They report that the asymmetry-threshold GARCH-MIDAS model can outperform the standard existing model significantly, but again the improvement is stronger in the case of short-term asymmetry and extreme volatility effects than the long-term effects. In the context of commodity price volatility forecasts, Pan *et al.* (2017) propose a Regime-Switching GARCH-MIDAS model for oil price volatility and report that the model outperforms the single-regime model in forecasting oil volatility.

With the development of more flexible and data intensive statistical methods in recent years, supervised machine learning (ML) models have also emerged as promising tools for volatility forecasting across different asset classes; see the recent review by Gunnarsson *et al.* (2024). As discussed in Christensen *et al.* (2023), machine learning techniques have inherent capabilities to deal with high-dimensional predictors with complex inter-dependencies. Given the large number of factors that can affect volatility of asset prices as well as differences in their frequencies, a natural approach is to combine GARCH models with ML techniques for modeling and predicting volatility. Therefore, the aim of this paper is to use a novel hybrid ensemble approach to forecast volatility of shipping freight rates. To this end, we investigate the performance of a variety of ML algorithms, GARCH, and GARCH-MIDAS models in forecasting volatility of shipping freight rates over short and medium horizons. We evaluate the forecasting accuracy gains by means of two statistical approaches: the Superior Predictive Ability (SPA) test of Hansen (2005) and the Model Confidence Set (MCS) method of Hansen *et al.* (2011). The SPA test focuses on the predictive ability of a predefined benchmark model with respect to several alternatives. It is widely applied to assess if specific assumptions for the multivariate structure, such as constant correlations, or the dynamics of individual volatility, such as short memory, are valid. With the MCS method, we identify from the initial set of competing models those which display equal predictive ability and outperform the others at a given confidence level. Both tests are executed using several symmetric and asymmetric loss functions, robust to the choice of the volatility proxy (Laurent *et al.* 2011, Boudt *et al.* 2013, Patton 2011).

This paper contributes to the literature in several ways. First, we conduct an extensive forecasting comparison between a variety of machine learning algorithms and econometric models of the ARCH family in forecasting volatility of spot freight rates over different horizons. We provide evidence of when and why some of these methods improve the accuracy of forecasting volatility. Second, we investigate the impact of several macroeconomic factors and market variables with different frequencies as predictors of dry bulk freight volatility. Third, we propose a novel market tightness index, capturing physical constraints in shipping markets, and use this index as a determinant of freight market volatility. Finally, we forecast the volatility of spot freight rates for dry bulk carriers of different sizes, used for physical trading, as well as of the underlying forward freight agreement contracts. We then compare the accuracy of volatility forecasts and the effects of exogenous factors on volatility, across vessel sizes and forecasts horizons.

The rest of this paper is structured as follows. Section 2 provides a review of the literature review on shipping volatility modeling and forecasting. In section 3 we provide details regarding the models we include. Section 4 describes the data collection and processing, offering insights about the frequency conversion and model validation approaches used. Section 5 presents the results of the extensive out-of-sample forecasting exercise. Finally, section 6 concludes.

## 2. Literature review

There is a large body of literature on modeling and forecasting shipping freight rates and asset prices. The first study on modeling volatility of shipping freight rates is Kavussanos (1997), where a GARCH model is applied to capture the volatility dynamics of spot and time-charter rates in the dry bulk shipping market. Early studies on modeling shipping market volatility concentrated on applications to various shipping assets and cross-sector comparisons (Kavussanos 1996, 1997). Later work on shipping market volatility concentrated on model specification and the effects of exogenous variables on behavior of volatility. Alizadeh and Nomikos (2011) investigate the relationship between the dynamics of the term structure and volatility of shipping freight rates. They argue, that due to the differences in elasticities of shipping supply curve and the shape/slope of forward curve—explained as the difference between short- and long-term freight rates—volatility of freight rate tends to increase when the market in backwardation and decrease when the market is in contango. Xu *et al.* (2011) conducted an investigation into the relationship between the time-varying

volatility of dry bulk freight rates and the changes in the supply of the fleet. They report that a change in fleet size positively affects the volatility of freight rates for larger vessels more than for smaller ones. Drobetz *et al.* (2012) investigate volatility dynamics in the dry bulk and tanker freight markets using different GARCH-X and EGARCH-X models and macroeconomic factors. They report that macroeconomic factors exhibit some explanatory power on freight rate volatility and such effects are more observed in the tanker market than in the dry bulk one. Xu *et al.* (2022) investigate the effects of COVID-19 on the Baltic Dry Index (BDI) volatility using a GARCH-MIDAS approach incorporating freight rates, Brent crude oil prices, container idle rates, port congestion levels, and global port calls as exogenous variables. They report that the increase in COVID-19 infection numbers impacted the BDI volatility regardless of the influence of other factors.

Another set of studies utilizes bivariate-GARCH models to estimate the volatility of spot and futures/forward freight rates and determine time-varying hedge ratios. For instance, Kavussanos and Nomikos (2000) and Kavussanos and Visvikis (2004) provide evidence that the time-varying hedge ratio determined by a bivariate GARCH-X model is more appropriate than the constant hedge ratio in terms of hedging performance when using Forward Freight Agreements (FFAs) for hedging dry-bulk shipping freight rates. Alizadeh *et al.* (2015) extend the Bivariate GARCH model to a Bivariate Markov Regime-Switching GARCH model to estimate the volatility of spot and forward tanker freight rates and assess the effectiveness of hedging tanker freight rates using forward freight agreements under different market conditions. More recently, Alizadeh and Sun (2023) propose a Conditional VaR model for the determination of hedge ratio for hedging dry bulk freight rates using FFAs and compare their performance with conventional GARCH and Regime Switching GARCH models.

More recent studies focus on forecasting freight rate volatility and VaR estimation using variety of GARCH models. For example, Abouarghoub *et al.* (2014) propose a regime-switching GARCH model for forecasting the volatility of tanker freight rates and the estimation of daily VaR. They report that the two-state MRS-GARCH performs better than the single regime model when subjected to the back-testing exercise. Angelidis and Skiadopoulos (2008) utilize several different specifications of GARCH models to estimate dry bulk and tanker freight rate volatilities and compare their performance against non-parametric methods in Value-at-Risk estimation using a back-testing approach. They find that the simplest non-parametric methods should be used to measure freight rate risk. Gavriilidis *et al.* (2018) examine whether inclusion of oil price shocks of different origins as exogenous variables in GARCH-X models improves the accuracy of their volatility forecasts for monthly spot and time-charter tanker earnings. They introduce exogenous variables based on three distinct oil price shocks—oil supply shock, aggregate demand shock, and precautionary oil-specific demand shock—distinguished through a VAR model, and integrate these variables in a GARCH-X specification. Their results suggest that aggregate oil demand shocks can improve the accuracy of freight rate volatility forecasts. Argyropoulos and Panopoulou (2018) compare the performance of different models including non-parametric historical simulation, GARCH, and combination forecasts in the estimation of daily VaR for tanker and dry bulk freight rates. However, they only compare one day ahead forecasts and VaR estimates and do not consider the effect of any market variables.

In a recent study, Liu *et al.* (2022) propose a support-vector regression GARCH (AR-SVR-GARCH) and asymmetric SVR-GJR GARCH models, which combine traditional time series analysis and modern machine learning methods to predict the volatility of the dry (BDI) and tanker (BDTI) shipping indices as well as of a shipping stock index (DJGS). They investigate the performance of these models in forecasting volatility of shipping freight rates using MSE and QLIKE statistical criteria and model confidence tests. While they report that both symmetric and asymmetric models of dry and tanker freight index volatilities are less affected by long-term trends, the volatility of the shipping stock index seems to be more affected by long-term trends. In addition, they report that the SVR-GARCH and the SVR-GJR models perform better in forecasting volatilities during periods of financial crisis and the recent turbulent shipping markets.

Overall, the review of the past studies suggests that volatility of shipping freight rates is time-varying, and the dynamics of the volatility can be explained by some macroeconomic and exogenous factors as well as market conditions. However, there is no proper investigation into the performance of ML and GARCH-MIDAS models in forecasting short- and long- run freight rate volatility. Moreover, the majority of current studies focus on the application of one or two machine learning algorithms. In contrast, this paper presents a thorough analysis of the out-of-sample performance of various tree-based algorithms (Random Forest, XGBoost, and CatBoost). Among the class of traditional econometrics models, we include the GARCH-MIDAS which, thanks to a component approach to volatility, is able to include macroeconomic and industry variables at lower frequency in the prediction of the long-run volatility.

## 3. Methodology

In this section, we explain the different approaches used in this paper to produce forecasts of freight rate volatility and evaluate their performance. Among supervised machine learning algorithms, both tree-based models and recurrent neural networks are capable of learning complex, non-linear time-series dynamics. It is well known that tree-based methods lend themselves to economic interpretation and are easier to apply in real-world applications. Hence, we focus on a representative set of tree-based ensemble models, more precisely the Random Forest, the XGBoost, and the CatBoost. These algorithms have been shown to perform well in relatively small samples and handle data disaggregation well. We don't include LightGBM among these due to the small sample size. Recurrent neural networks and deep learning models, such as LSTM, require a much higher number of data points to accurately distinguish signal from noise. We

have only 1473 observations. Furthermore, deep neural networks offer limited interpretability due to their black-box nature and require careful hyperparameter tuning and training to provide reliable results. Finally, in the case of mixed frequency data, as in our sample, the temporal disaggregation approaches may cause significant noise and thus prevent the algorithm from correcting learning data patterns. On the other hand, tree-based ensemble methods have been shown to have good performance on disaggregated data in feature engineering forecasting. Furthermore, they allow quite direct interpretations and comparison with the GARCH models.

The freight rate log return $r_t$ at time $t$ is modeled as

$$r_t = \mu_t + \epsilon_t, \quad \epsilon_t = \sigma_t z_t \tag{1}$$

and our aim is to forecast its daily conditional volatility, defined as

$$\sigma_t^2 = Var[r_t \mid F_{t-1}] = E[(r_t - \mu_t)^2 \mid F_{t-1}] \tag{2}$$

where $F_{t-1}$ represents the sigma algebra of all the information available at time $t-1$ and $z_t$ is an i.i.d. shock with distribution D. Since volatility itself is latent thus unobservable, we must choose a proxy for it. Andersen and Bollerslev (1998) showed that the demeaned squared daily returns are an unbiased estimator of the volatility, albeit quite noisy. A significantly more accurate measure of volatility is the realized variance (RV) calculated from intra-day prices, however data on freight rates is available only at daily frequency; thus, in this study we use the demeaned squared daily returns as a target variable when training the machine learning algorithms and as true ex-post volatility in the forecasting comparison.

### 3.1. The GARCH model and its extensions

Since the seminal work of Engle (1982), an extensive body of literature on modeling the temporal dependencies in financial market volatility using the discrete GARCH model has emerged. Given the parsimonious nature of the GARCH model, the intricate structure of the underlying data renders it inadequate in certain aspects. After its inception, a large number of extensions of the basic specification have been developed. These extensions include the GJR-GARCH model of Glosten *et al.* (1993) and the EGARCH Nelson (1991), the component GARCH model of Engle and Lee (1999) and many more specifications (see Franq and Zakoian 2019 for an extensive overview). These models have been extensively used in the investigation of the dynamics of dry bulk market volatility, as discussed in section 2. In this framework, the conditional variance is expressed as a deterministic function of past returns. The GARCH($p$, $q$) model is defined as:

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{t-1}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-1}^2, \tag{3}$$

with constraints $\beta_j \geq 0$ and $\sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) < 1$ to ensure respectively positiveness and stationarity of the conditional variance. To ensure our results are robust against a broader suite of recent GARCH-type models, we include several of the more popular ones in the forecasting comparison. Table 1 summarizes the GARCH-specifications utilized in this paper. We test the GJR and EGARCH specifications to allow for asymmetric responses of the volatility to positive and negative shocks. We include the FIGARCH model to allow for fractional integration in the volatility decay rate. All these extensions are still based on past returns only as conditional information. Recently, however, several studies recognized the relationship between volatility and the macroeconomic factors. In particular, Engle *et al.* (2013) propose a GARCH model which utilizes exogenous mixed data sampling known as GARCH-MIDAS to incorporate the effect of macroeconomic factors on the dynamics of volatility including short- and long-term components. In the GARCH-MIDAS model, a short-run variance GARCH component fluctuates around a time-varying long-term component that is a function of macroeconomic or financial explanatory variables. By allowing for a mixed-frequency setting, this approach bridges the gap between daily stock returns and low-frequency (e.g. monthly, quarterly) explanatory variables. In the GARCH-MIDAS specification, the volatility is specified as

$$\sigma_t^2 = \tau_t \times g_t \zeta_t \quad \zeta_t \sim D(0,1) \tag{4}$$

where $\tau_t$ and $g_t$ are respectively the long-term and short-term volatility components. The short-run component adheres to a straightforward mean-reverting GARCH(1,1) process, and fluctuates around the time-varying long-run volatility. The long-run volatility component captures via a MIDAS filter the impact of the exogenous features $X_{t,j}$ available at lower frequencies:

$$\tau_t = \exp\left( m + \theta_j \sum_{k=1}^{K_j} \delta_{k,j}(\omega) X_{t-k,j} \right) \quad \forall j \in J \tag{5}$$

where $m$ represents a constant term, and $\delta_{k,j}(\omega)$ is the MIDAS weight function (Engle *et al.* 2013). We test the Beta and Exponential Almon lag functions (5). In this specification, we test as predictor of the long-run volatility the Market Tightness Index, which we introduce in section 4.

### 3.2. The random forest

The Random Forest (RF) has become a prominent and widely-used machine learning algorithm. Random forests' basic philosophy is based on combining three concepts: (i) classification or regression decisions trees, (ii) bootstrap aggregation or bagging and (iii) random subspaces. As a significant implementation of the bagging framework, the Random Forest generates a large number of de-correlated trees and then combines them to create an ensemble prediction, and thus, improving overall accuracy by reducing the variance by averaging the noisy and unbiased trees. Its structure follows a divide-and-conquer approach used to capture nonlinearity in the data and perform pattern recognition. The algorithm's name suggests a collection of diverse trees, like a forest, varying in shape and size. For a given data set, a continuous outcome variable and a set of features and a number of trees, the outline of the algorithm is as follows: (i) generate a bootstrapped dataset

Table 1. The univariate volatility models applied this paper, with corresponding formulas and parameters.

| Model | Specification | Parameters |
|---|---|---|
| GARCH(1,1) | $\sigma_t^2 = \alpha_0 + \alpha_1 a_{t-1}^2 + \beta_1 \sigma_{t-1}^2$ | $\alpha_0, \alpha_1, \beta_1$ |
| EGARCH(1,1) | $\ln\left(\sigma_t^2\right) = \alpha_0 + \alpha_1 \left\lvert \frac{a_{t-1}}{\sigma_{t-1}} \right\rvert + \gamma_1 \frac{a_{t-1}}{\sigma_{t-1}} + \beta_1 \ln\left(\sigma_{t-1}^2\right)$ | $\alpha_0, \alpha_1, \beta_1, \gamma_1$ |
| GJR-GARCH(1,1) | $\sigma_t^2 = \alpha_0 + (\alpha_1 + \gamma_1 N_{t-1}) a_{t-1}^2 + \beta_1 \sigma_{t-1}^2$ | $\alpha_0, \alpha_1, \beta_1, \gamma_1$ |
| FIGARCH(1,d,1) | $\sigma_t^2 = \alpha_0 [1 - \beta(L)]^{-1} + \{1 - [1 - \beta(L)]^{-1} \phi(L)(1 - L)^d\} a_{t-1}^2$ | $\alpha_0, \beta, d$ |
| GARCH(1,1)-MIDAS | $\sigma_t^2 = g_t \times \tau_t$ | |
| | $= \left[(1 - \alpha - \beta) + \alpha \frac{(r_t - \mu_t)^2}{\tau_t} + \beta g_{t-1}\right]$ | |
| | $\times \left[\exp(m + \theta_j \sum_{k=1}^{K_j} \delta_{kj}(\omega) X_{t-k,j})\right], \quad \forall j \in J$ | $\alpha, \beta, m, \theta_j, \omega$ |

from the initial dataset, (ii) use the bootstrap sample to grow a tree. At each node, perform the following steps: (a) randomly select $d$ features without replacement. (b) Segment the node by choosing the feature that provides the best split according to the objective function. (iii) Repeat steps 1 and 2 for each tree under consideration. Each tree in this approach is identically distributed, so the expected value of averaging $N$ trees is the same as any single tree, implying that the bias of the bagged trees is the same as that of the individual bootstrap trees and variance reduction is targeted for forecasting accuracy. For construction, the predictive ability of RFs increases as the inter-tree correlation decreases. Thus, a large number of predictors can provide increased generalization capacity, by having each tree randomly select a number $m$ splitting candidates from $p$ variables, such that $m \leq p$ to minimize the trees' correlation. It is critical for the overall performance of random forests to find the optimal of $m$ and the optimal number of trees via hyperparameter selection. A detailed discussion about hyperparameter and feature selection can be found in section 4.

### *3.3. XGBoost*

The combination of decision trees and gradient boosting methods has the advantages of good training effect and not easily over-fitting. Gradient boosting is a generalization of tree boosting designed to address various issues with regular boosting, namely speed, interpretability, and, in some cases, robustness against overlapping class distributions. The XGBoost, developed by Chen and Guestrin (2016), is an ensemble model which consists of an efficient implementation of decision trees, in order to produce a combined model whose predictive performance is better than individual techniques used alone. Differently from bagging, however, boosting does not carry out bootstrap sampling, but trees are grown in a sequential basis entailing that the current generated tree exploits information from the previously generated tree. Hence, trees are no longer grown independently but sequentially dependent on construction. The additive aspect of the algorithm embodies the core principle of boosting, which iteratively adds trees to reduce the loss incrementally. This involves parameterizing each tree and adjusting these parameters to minimize the residual loss. The output of each newly added tree is then combined with the outputs of previously added trees to improve the model's overall performance. This process continues, adding a predefined number of trees until training stops, either when the loss reaches an adequate threshold or when validation loss converges (Brownlee 2017). A tree ensemble model utilizes $K$ additive functions to forecast the result, where $T$ denotes the number of leaves in each tree. Each function $f_k$ represents an independent tree structure $q$ and leaf weights $w$. Regression trees assign a continuous score to each leaf, represented by $w_i$ for the $i$th leaf. Further, the decision rules specified by $q$ in the trees determine the leaf to which any example x is assigned. The final prediction is derived by summing the scores of the relevant leaves, denoted by $w$. Consequently, the mathematical formulation of the additive model begins by introducing the regularized objective to be minimized as in (6):

$$\mathcal{L}(\phi) = \sum_i l\left(\hat{y}_i, y_i\right) + \sum_k \Omega\left(f_k\right), \quad \text{where}$$
$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 \tag{6}$$

where $l$ represents the loss function and $\Omega$ denotes the regularization term, which penalizes model complexity. One of the key advantages of gradient boosting is its flexibility to accommodate various differentiable loss functions within a single boosting framework. Given that squared residuals are both the default loss function and well-suited for numerical values, we use the mean squared error (MSE) as the loss function. Incorporating the MSE loss function results in a simplified expression, which encompasses both a quadratic and first-order residual term:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} \left(y_i - \left(\hat{y}_i^{(t-1)} + f_t\left(x_i\right)\right)\right)^2 + \sum_{i=1}^{t} \Omega\left(f_i\right) \tag{7}$$

In the XGBoost, several parameters need to be tuned to maximize the power of model performance and to prevent overftting problems, including the number of trees, the number of tree splits, the learning rate, the number of iterations, the maximum depth. A detailed discussion about hyperparameter tuning and feature selection can be found in section 4.

### *3.4. CatBoost*

The categorical boosting algorithm, CatBoost was introduced by Gulin *et al.* (2018). The algorithm derives the first part of its name from categorical features. Unlike traditional gradient boosting that converts these features to numbers before

training, CatBoost processes them *during* training. Another appealing attribute is its efficient strategy to mitigate overfitting while using the entire dataset for training. Specifically, the algorithm performs a random permutation of the dataset. For each example in the dataset, the average label value is derived from preceding examples in the permutation that share the identical category value. This often results in better predictive performance than XGBoost and other gradient-boosted tree algorithms (Gulin *et al.* 2018). The permutation is dented by $\sigma = (\sigma_1, \ldots, \sigma_n)$. To compute the transformed value for each example, the algorithm relies on:

$$\frac{\sum_{j=1}^{p-1} \left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] Y_{\sigma_j} + \alpha \cdot P}{\sum_{j=1}^{p-1} \left[x_{\sigma_j,k} = x_{\sigma_p,k}\right] + \alpha} \tag{8}$$

where $\left[x_{\sigma_j,k} = x_{\sigma_p,k}\right]$ is an indicator function that equals 1 if the category values $x_{\sigma_j,k}$ and $x_{\sigma_p,k}$ match, and 0 otherwise. The prior value is expressed by $P$ with a corresponding parameter $\alpha > 0$, which determines the weight of the prior. Employing a prior is a standard method to minimize noise from low-frequency categories. In a regression framework, the prior is typically calculated as the average label value in the dataset. CatBoost introduces an innovative approach for calculating leaf values, enabling multiple permutations without the risk of overfitting. CatBoost leverages oblivious trees as its base predictors, utilizing a consistent splitting criterion at each level to maintain tree balance and minimize overfitting: all features are transformed into a binary format to optimize prediction accuracy. This binary encoding method allows for efficient calculation of leaf indices, resulting in quicker and more precise model predictions. Furthermore, the entire computation process can be executed in parallel, achieving up to a threefold increase in speed, making the model exceptionally efficient (Gulin *et al.* 2018). The tree structure in CatBoost is chosen through a greedy method. Features and their corresponding splits are sequentially selected for substitution in each leaf. The selection of candidates is derived from the initial split calculations and the conversion of categorical features into numerical features. The tree depth and other structural rules are determined by the initial parameters. The approach for choosing a feature-split pair for a leaf involves several steps. Initially, a list of potential candidates (feature-split pairs) is created to be considered for assignment to a leaf. Subsequently, penalty functions are calculated for each candidate, assuming they have all been allocated to the leaf. Then, the split with the least penalty is chosen. Finally, this selected value is allocated to the leaf. This process is repeated for all subsequent leaves, ensuring that the number of leaves corresponds to the tree's depth. While the literature on CatBoost primarily highlights its ability to handle categorical features, we rely on this method for its use of oblivious trees. This enables us to evaluate the forecasting performance of a boosting model with a lower risk of overfitting compared to models using more complex tree structures.

## 4. Data description and processing

This paper uses a sample of spot freight rates for the four main dry bulk shipping segments: Capesize, Panamax, Supramax, and Handysize vessels from the Baltic Exchange, along with industry variables such as fleet size, sales, age, seaborne trade in commodities, and fuel prices from Clarksons' Shipping Intelligence Network(SIN)† All freight rates are expressed in USD/day and reflect the average daily hire rates of the corresponding vessel size on that day. The sample consists of daily data and spans from 1 of November 2017 to 28 of September 2023, for a total of 1473 observations. It is worth noting that due to the nature of the market, reliable and consistent published data on dry bulk freight rates are only available on a daily frequency. The industry variables cover the same period and are measured at a monthly frequency.‡

The Baltic Exchange average trip-charter freight rate for different size dry bulk carriers is constructed in a composite manner, aiming to represent the trading activities of each vessel type in major routes globally and to capture the observed trade flows in the markets. The weighted averages of trip-charter rates are labeled based on the typical vessel type and size (see table A1 in Appendix 1 for the composition of average trip-charter rates for each vessel type). These average trip-charter rates reflect the spot market level for each vessel type on any day and used for settlements of forward freight agreements and freight options traded for maturities from one month to several years. The average Trip-Charter rates are also used by market participants to benchmark their operational efficiency as well as negotiations and physical shipping contracts on a floating freight rates basis. For instance, a commodity trader can hire a vessel from a shipowner for one year at the Baltic Capesize Average 5TC plus 5%. Thus, information about the level and behavior of volatility dynamics of these freight rates is of utmost importance for shipowners, operators, and traders alike. Figure 1 presents average trip-charter rates for different size dry bulk carriers over the sample period. It can be seen that shipping freight rates can fluctuate significantly over short periods. The noticeable drop in freight rates for all vessel sizes is during the early stages of COVID-19 pandemic (2020) and a sharp recovery after easing of the lock down measures around the world (2021) when the economy went through a V shape recovery. There is also a sharp increase due to the war in Ukraine in the first half of 2022, and easing of the freight market possibly due to the higher inflation rates combined with mild economic recession around the world. However, there are also significant shorter term movements across all vessel sizes due to other random shocks.

Descriptive statistics of the log-return of freight series for different vessel types are reported in table 2. The annualized average daily returns are positive and increase with vessel size, indicating a general increase in freight rates

† Freight rates are defined as average trip-charter rates for each vessel type defined by the Baltic Exchange. Capesize, Panamax, Supramax, and handysize vessels are defined as a 180 000dwt, 82 000dwt, 58 000dwt, and 38 000dwt bulk carrier, respectively. The average trip-charter rates, displayed in figure 1, are: Average 5TC for Capesize, Average 5TC for Panamax, Average 10TC for Supramax, and average 7TC for Handysize ships, respectively. See the Baltic Exchange website https://www.balticexchange.com/en/data-services/market-information0/indices.html for further details on average trip-charter rates.

‡ The list of industry variables as proxies for supply and demand for the dry bulk shipping sector is listed in table 3.

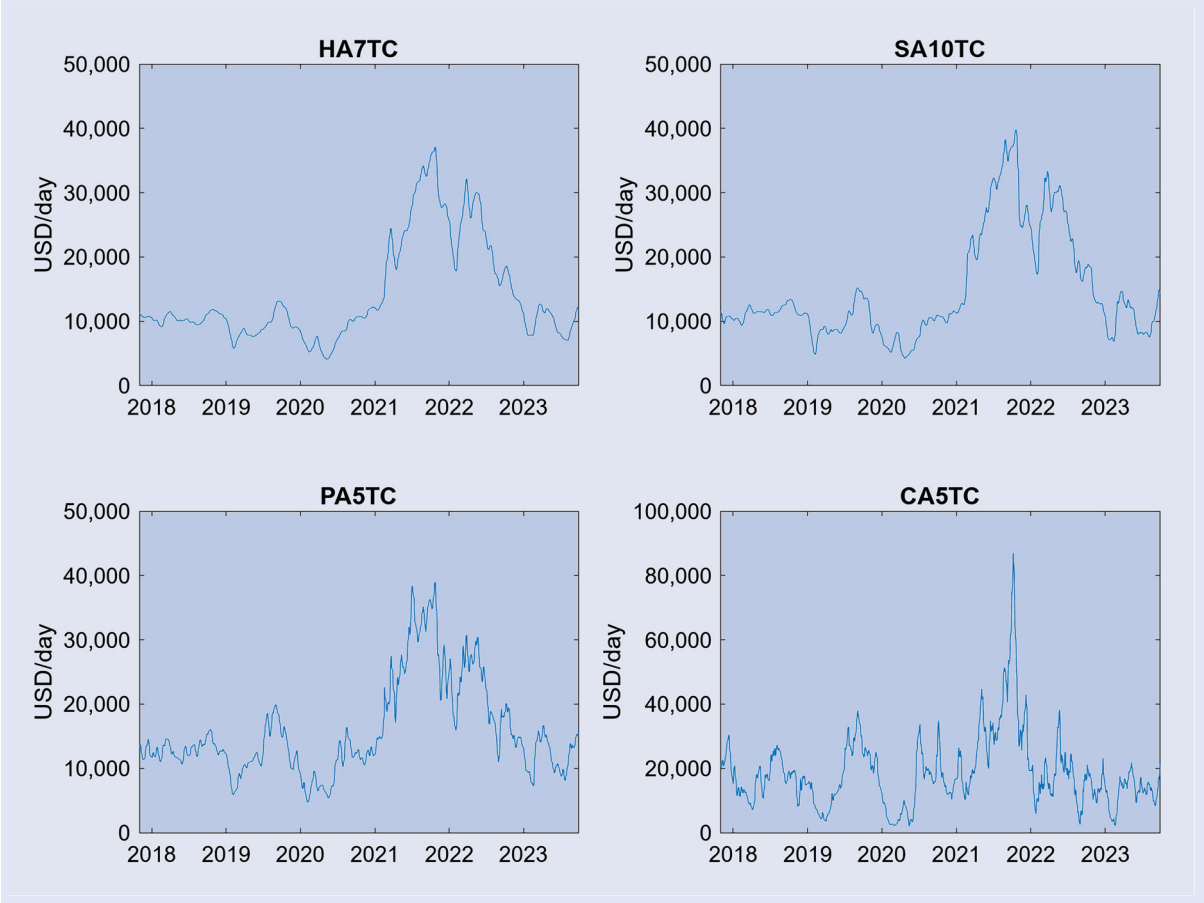


Figure 1. Baltic exchange average trip-charter rates.

Table 2. Descriptive statistics for daily log returns of Baltic Exchange average trip-charter rates.

| Series | $T$ | Min | Max | Mean | Dev. | Std. Skew. | Excess Kurt. | JB Statistic | JB $p$-value |
|---|---|---|---|---|---|---|---|---|---|
| HA7TC | 1472 | − 7.75% | 6.52% | 3.32% | 20.7% | − 0.1158 | 5.3234 | 316.20 | < 0.0001 |
| SA10TC | 1472 | − 9.56% | 10.41% | 4.24% | 29.0% | 0.1719 | 7.7143 | 1295.87 | < 0.0001 |
| PA5TC | 1472 | − 14.81% | 24.05% | 2.74% | 48.0% | 0.6944 | 7.8298 | 1464.81 | < 0.0001 |
| CA5TC | 1472 | − 36.03% | 44.67% | 9.26% | 119.7% | 0.6780 | 8.4326 | 1818.41 | < 0.0001 |

| | Level series | | | | Log return series | | | |
|---|---|---|---|---|---|---|---|---|
| | ADF Test | | PP Test | | ADF Test | | PP Test | |
| | Statistic | $p$-value | Statistic | $p$-value | Statistic | $p$-value | Statistic | $p$-value |
| HA7TC | − 1.7511 | 0.7279 | − 0.9574 | 0.9496 | − 8.2519 | < 0.0001 | − 9.2789 | < 0.0001 |
| SA10TC | − 1.8591 | 0.6756 | − 1.3446 | 0.8765 | − 10.6433 | < 0.0001 | − 9.9621 | < 0.0001 |
| PA5TC | − 2.3043 | 0.4316 | − 2.2216 | 0.4777 | − 14.4219 | < 0.0001 | − 13.0910 | < 0.0001 |
| CA5TC | − 4.1756 | 0.0049 | − 3.3112 | 0.0644 | − 13.0530 | < 0.0001 | − 19.5221 | < 0.0001 |

Note: Sample period is from 1 of November 2017 to 28 of September 2023.
Summary statistics include the minimum, maximum, mean, standard deviation (Std.Dev.), skewness, excess kurtosis, Jarque-Bera (JB) statistic, and the associated p-value. The ADF and PP tests are the Augmented Dickey–Fuller and the Phillips–Perron Unit root tests, respectively.

over the sample period. Similarly, annualized standard deviations indicate significant unconditional volatility which is directly related to vessel sizes. The high level of standard deviation for Capesize freight rates (119.7%) is a clear evidence of the highly volatile nature of this sector. The coefficients of skewness and kurtosis reveal that the return series are skewed with a high degree of excess kurtosis, and the results of the Jarque–Bera test statistics strongly indicate that the returns are not normally distributed. Furthermore, the results of the Augmented Dickey-Fuller and Philips and Perron unit root tests, reported in the lower panel of table 2, suggest that freight rate series across all vessel sizes are non-stationary in levels, but stationary in first-order log differences.

Table 3. List of industry and macro variables as exogenous predictors.

| No | Series | Supply/demand | Frequency | Mean | St Dev |
|---|---|---|---|---|---|
| 1 | Handysize fleet size (dwt) | S | Monthly | 109.41 | 4.74 |
| 2 | Supramax fleet size (dwt) | S | Monthly | 212.65 | 11.58 |
| 3 | Panamax fleet size (dwt) | S | Monthly | 225.34 | 15.87 |
| 4 | Capesize fleet size (dwt) | S | Monthly | 358.19 | 21.48 |
| 5 | Handysize orderbook (dwt) | S | Monthly | 8.74 | 2.07 |
| 6 | Supramax orderbook (dwt) | S | Monthly | 17.97 | 3.29 |
| 7 | Panamax orderbook (dwt) | S | Monthly | 22.45 | 3.33 |
| 8 | Capesize orderbook (dwt) | S | Monthly | 36.36 | 11.74 |
| 9 | Handysize demolition (dwt) | S | Monthly | 0.051 | 0.048 |
| 10 | Supramax demolition (dwt) | S | Monthly | 0.079 | 0.081 |
| 11 | Panamax demolition (dwt) | S | Monthly | 0.076 | 0.104 |
| 12 | Capesize demolition (dwt) | S | Monthly | 0.410 | 0.472 |
| 13 | Handysize Average Speed (knots) | S | Daily | 11.16 | 0.16 |
| 14 | Supramax Average Speed (knots) | S | Daily | 11.33 | 10.7 |
| 15 | Panamax Average Speed (knots) | S | Daily | 11.31 | 0.16 |
| 16 | Capesize Average Speed (knots) | S | Daily | 11.16 | 0.24 |
| 17 | Fuel oil price (USD/mt) Singapore HSFO 380cst | S | Weekly | 416.13 | 102.27 |
| 18 | Handysize Bulker Sales (1000 dwt) | S | Monthly | 499.03 | 257.12 |
| 19 | Suoramax Bulker Sales (1000 dwt) | S | Monthly | 1104.19 | 477.20 |
| 20 | Panamax Bulker Sales (1000 dwt) | S | Monthly | 4011.62 | 503.88 |
| 21 | Capesize Sales (1000 dwt) | S | Monthly | 1119.67 | 818.43 |
| 22 | Handysize Bulkcarrier Fleet - Average Age | S | Monthly | 11.66 | 0.86 |
| 23 | Supramax Bulkcarrier Fleet - Average Age | S | Monthly | 10.10 | 0.98 |
| 24 | Panamax Bulkcarrier Fleet - Average Age | S | Monthly | 10.28 | 0.85 |
| 25 | Capesize Bulkcarrier Fleet - Average Age | S | Monthly | 8.88 | 0.67 |
| 26 | World Steel Production (1000) | D | Monthly | 153.89 | 8.99 |
| 27 | Global Seaborne Iron Ore Trade Indicator (Volume Index) | D | Monthly | 115.5 | 8.5 |
| 28 | Global Seaborne Coal Trade Indicator (Volume Index) | D | Monthly | 104.3 | 7.5 |
| 29 | Global Seaborne Grain Trade Indicator (Volume Index) | D | Monthly | 126.95 | 12.6 |
| 30 | Global Seaborne Minor Bulk Trade Indicator (Volume Index) | D | Monthly | 118.8 | 5.4 |
| 31 | Global Seaborne Dry Bulk Trade Indicator (Volume Index) | D | Monthly | 112.32 | 5.39 |
| 32 | Global Seaborne Iron Ore Trade Indicator (% Yr/Yr) | D | Monthly | 0.65 | 5.82 |
| 33 | Global Seaborne Iron Ore Trade Indicator (% Yr/Yr 3mma) | D | Monthly | 0.7 | 3.74 |
| 34 | Global Seaborne Coal Trade Indicator (% Yr/Yr) | D | Monthly | 1.52 | 9.08 |
| 35 | Global Seaborne Coal Trade Indicator (% Yr/Yr 3mma) | D | Monthly | 1.67 | 771 |
| 36 | Global Seaborne Grain Trade Indicator (% Yr/Yr) | D | Monthly | 1.87 | 8.58 |
| 37 | Global Seaborne Grain Trade Indicator (% Yr/Yr 3mma) | D | Monthly | 1.98 | 6.69 |
| 38 | Global Seaborne Minor Bulk Trade Indicator (% Yr/Yr) | D | Monthly | 1.54 | 5.67 |
| 39 | Global Seaborne Minor Bulk Trade Indicator (% Yr/Yr 3mma) | D | Monthly | 1.52 | 4.63 |
| 40 | Global Seaborne Dry Bulk Trade Indicator (% Yr/Yr) | D | Monthly | 1.10 | 3.79 |
| 41 | Global Seaborne Dry Bulk Trade Indicator (% Yr/Yr 3mma) | D | Monthly | 1.17 | 2.85 |
| 42 | OC Production - Total Industry Excl. Construction SADJ | D | Monthly | 103.75 | 3.84 |
| 43 | OC Production - Total Industry Excl. Construction VOLA | D | Monthly | 0.72 | 5.68 |

Note: Sample period is from 1 of November 2017 to 28 of September 2023.
S denotes supply-specific, while D represents demand-specific variables. The variables are measured in different frequencies and some are transformed to first-order log differences. All variables are stationary according to the ADF test at significance level of $\alpha = 5\%$.

We split the data into a training, validation, and test set. The last year (250 days) of data is the test set, reserved for out-of-sample evaluation. We split the remaining dataset in a training set of 90% of the dataset, a validation set of 10%. We conduct a robustness check, where the training set is modified to 80% and with a validation set fixed at 20%. The results are broadly in line with our reported findings.

Following the findings of previous studies on the effect of macroeconomic and industry variables on the volatility of freight rates (Xu *et al.* 2011, Drobetz *et al.* 2012), we collected several variables which relate to either the demand or the supply side of the shipping freight market. A full list of the explanatory variables, collected from Clarksons' Shipping Intelligence Network (SIN), as well as their acronyms and frequencies, is presented in table 3. The first set of variables is considered to affect the supply-side of shipping services and includes fleet size, sales volume, and vessel age in different sectors as well as fuel oil prices. For example, the fleet size for different types of vessels (in dwt), represents the fleet's overall capacity to transport cargo at any point in time, while sales volume indicates the willingness to invest in ships, and age of fleet can be considered as a proxy for fleet productivity and increase in supply due to recent new-building deliveries. Additionally, the age of a vessel can impact its operational efficiency; an ageing fleet suggests reduced efficiency, which could have a negative impact on supply.

On the demand side, the variable set includes seaborne trade volume in major bulk commodities such as iron ore, coal, grain, as well as minor bulk commodities (e.g. bauxite and alumina, phosphate rock, fertilizers, cement, rice, etc.),

Table 4. Descriptive statistics of market tightness index.

| Series | *N* | Min | Max | Mean | Std. dev | Skew. | Kurt. | JB Stat | JB *p*-val |
|---|---|---|---|---|---|---|---|---|---|
| MTP | 70 | 43.7937 | 176.2696 | 105.6996 | 25.2234 | 0.3028 | 1.3395 | 49 569 | 0.0544 |
| Change in MTP | 69 | − 0.3140 | 0.3910 | 0.0051 | 0.1257 | 0.2539 | 1.1800 | 3.6774 | 0.0890 |

Note: Summary statistics for Market Tightness Index including count (N), minimum, maximum, mean, median, standard deviation (Std. Dev.), skewness, excess kurtosis, Jarque-Bera (JB) statistic, and associated *p*-value.

and the industrial production - excluding construction - of the OECD zone. Although there might be a certain degree of correlation between these demand side variables, they may have effects on the volatility of freight rates in different shipping sectors. For instance, iron ore is mainly carried by Capesize vessels, while grain is mainly carried by Panamax and Supramax vessels, and minor bulk commodities are mainly transported by Handysize vessels. Furthermore, we consider the year-on-year change and the three-month moving averages of year-on-year changes. These transformations are selected to provide a more granular view of major bulk commodities as they directly measure the volume of cargo transported within the dry bulk sector, capturing demand dynamics. Finally, the OECD industrial production level reflects the current demand for raw materials, while its volatility is an indicator of economic uncertainty, both of which are key proxies for the demand for dry bulk shipping.

It is well documented that the shipping freight market is cyclical (Stopford 2009, Alizadeh and Nomikos 2009) and the behavior of volatility can depend on the phases of the shipping cycle, reflecting supply and demand conditions. To incorporate such information in our models, we introduce a Market Tightness Index (MTP) which summarizes the supply and demand balance in the dry bulk market at time $t$:

$$MTP_t = \left( \frac{\text{Total Seaborne Trade}[\text{'000 tonnes}]}{\text{Fleet Size}[\text{dwt million}]/\text{Fuel Oil Price}\left[\frac{USD}{\text{tonne}}\right]} \right) \tag{9}$$

The descriptive statistics of the level and logarithmic changes in the market tightness index are presented in table 4. The market tightness index is constructed in such a way that a high level of the index indicates a tight market condition as the demand for dry bulk cargo transportation is greater than the supply and any changes in demand can have a great impact on the freight level. Similarly, a low level index is an indication of low demand and excess supply, that is when changes in demand can be absorbed by excess supply, which result in low freight volatility. The index is tested as an exogenous predictor in the machine learning models and in the GARCH-MIDAS model as a driver of the long-run volatility component. In developing our model specifications, we examine an extended set of variables and we experiment with various transformations of these variables to derive potentially more representative drivers to train our models, specifically we include: (i) simple log transformations of the candidate predictors, (ii) quarter-over-quarter change (first difference) and the quarter-over-quarter percentage change (relative difference) for every predictor. We also include lagged values of the predictors up to five lags. This process leads to a set of almost 210 predictors as potential candidates for our modeling procedures. Our decision to explore various transformations and lags of all predictors is motivated by the lack of conclusive evidence in current literature on the number of lags to include and whether using levels or first differences is preferable in predicting freight rates volatility.

### *4.1. Temporal disaggregation*

Several of the exogenous features that we test as predictors are available at low-frequencies, typically at monthly opening and closing values. In contrast, the target variable is at higher frequency, with daily observations. In the GARCH-MIDAS model, the MIDAS filter allows the use of lower frequency indicators in the long-run volatility component equation, but in all the machine learning algorithms this frequency mismatch can cause severe bias. This is a frequent challenge for researchers and several effective methods have been proposed in the literature to address it (Sax and Steiner 2013). It is well known that listwise deletion, where all entries with missing values are removed before the analysis, is easy to implement but has significant disadvantages: missing information can introduce bias and loss of precision (Little and Rubin 2019). More sophisticated techniques involve imputation—the statistical process of replacing missing values (Moritz and Bartz-Beielstein 2017). For example, forward filling is a simple and popular approach that leverages the temporal structure by using the most recent available observation to replace missing values (Che *et al.* 2018, Lipton *et al.* 2016). In a time series context however, interpolation methods which use information from previous and future data points have been shown to be more effective in handling frequency mismatch (Junninen *et al.* 2004). As outlined in Sax and Steiner (2013), the goal of temporal disaggregation is to derive an unknown high-frequency series $x$ that aligns with the sums, averages, or specific values (*first* or *last*) of a known low-frequency series $x_l$. The process involves modeling the differences between the observed low-frequency series and a higher-frequency series which is used as a proxy in the disaggregation process (Sax and Steiner 2013). Disaggregation methods differ in how they identify the high-frequency proxy and the mapping via the distribution matrix (Sax and Steiner 2013). In this paper, we utilize the Denton-Cholette transformation approach, first converting the log differences of all monthly frequent exogenous features, including the market tightness index, as outlined in table 3. We employ a constant value of one as the initial indicator, facilitating temporal disaggregation without the need of high-frequency indicator series. In this way, we effectively perform an interpolation that adheres to the temporal additivity constraint (Sax and Steiner 2013). All the frequency-transformed Denton-Cholette variables and the exogenous variables available at

daily frequency are included in the first calibration step of the machine learning models with exogenous features. This results in a diverse set of potential exogenous features to be included in the models. Variable importance and selection during model calibration is discussed in the next subsection. Each exogenous variable is aligned with the date entries of the target volatility proxy in the dataset to create the main input dataframe. Feature date entries that do not match the date entries of the target variable are removed from the input dataframe.

### *4.2. Model development and validation*

Model selection is a crucial step in the machine learning and econometrics modeling. The objective of this process is to select the most suitable model from a range of candidates, using an appropriate error measure. For all the machine learning models introduced in Section 3, we consider two configurations: (i) a pure time series configuration, where the input dataframe consisting solely in lagged values of the volatility proxy, and (ii) a configuration with exogenous predictors where, in addition to all exogenous variables, including the Market Tightness Index, are tested. The choice of these two configurations enables us to discuss the role of exogenous predictors in freight rate volatility forecasting and draw conclusions robust to model specification.

Development of recent data analysis techniques such as flexible Machine Learning approaches not only allows handling of large amount of data but also exploration of complex and nonlinear relation among the variables which could be used to produce better forecasts. Such complexities cannot be utilized in traditional econometric and time series models, where economic relationships are established and examined through statistical tests. Therefore, while a shortcoming of the ML approach is that economic relationships between variables cannot be established and tested, the advantage of handling a large data set and allowing for complex nonlinear interrelation can enhance prediction and forecasting performance of ML models.

Shipping is a complex industry where four main physical markets, namely the freight, second-hand, new-build, and demolition markets, constantly interact, while several variables within each market, including ship values, fleet size, orderbook, and shipping freight rates, among others, evolve at different speeds and rates. In addition, changes in international seaborne trade as well as decisions and actions by participants and agents in different shipping sub-markets directly and indirectly affect each segment of the industry.† Such decisions affect the supply and demand for shipping services and consequently impact shipping freight rates in the short- and long-run, and the level and dynamics of volatility of shipping freight rates. Therefore, investigating the behavior and volatility of freight rates in isolation may not be optimal. Using a broad range of industry variables in conjunction with flexible machine learning techniques can provide models with the context and richness of data required to capture the multifaceted nature of volatility of shipping freight, which could improve the accuracy and reliability of forecasts.

In machine learning applications, K-fold cross-validation is commonly recognized as the standard validation method (Schnaubelt 2019b). However validation strategies that maintain the temporal order of observations between the training and the validation set are more effective in time series dataset which may exhibit long-range dependence (Bergmeir *et al.* 2018). These are known as forward-validation methods. We use the growing-window validation technique of Schnaubelt (2019a) with $K = 5$ folds with a 90/10 training-to-validation set split to ensure proper model evaluation for future selection and hyperparameter tuning. For the machine learning models incorporating exogenous features, feature selection and hyperparameter tuning are conducted in two steps. First, we identify a subset of relevant features from the high-dimensional dataset based on specific evaluation criteria to reduce computational complexity and enhance the generalization ability of the models (Chen and Chen 2015). At this stage, we include all the exogenous variables, their transformations and lagged values and the lagged values of the volatility proxy as features, up to a predefined maximum number of lags (set to five for all models). We trim the initial large set of features using the feature importance scores on a the basic leaner specification which includes all features at our disposal (Wujek *et al.* 2016). The importance scores threshold is optimized by means of a specialized *wrapper* scheme which ensures that a feature subset selection algorithm acts as a wrapper around the induction algorithm (our simple learner), using the induction algorithm to evaluate different subsets of features. The subset with the highest evaluation is selected for the final model (Kohavi and John 1997). We use the list of feature importances from the initial model as candidate thresholds. This list is used to filter subsets of features whose importance is greater than or equal to the current threshold. These selected features are used to train a new model with a simple structure. Finally, we employ a forward selection approach which starts with an empty set of features and iteratively adds one feature at a time to the induction algorithm until all features are included. A valid subset of features must contain at least two features, and the iteration list of feature importance includes only one value of zero to keep the list as concise as possible. To evaluate the performance of a candidate threshold, we use a five-fold forward-validation. The performance of each candidate threshold is assessed based on the Mean Squared Error (MSE) in the validation set for each fold. We compute a weighted average of these MSE values with weights corresponding to the fold numbers, such that the fold containing the most training data receives the highest weight. The features which correspond to the best-performing threshold are then used in the subsequent hyperparameter tuning process. Figure 2 illustrates the feature selection process described above.

In the tuning process, we use a straightforward grid search strategy for all models, creating a grid of possible hyperparameter values from all combinations of predefined sets (Wujek *et al.* 2016). Appendix 2 contains the hyperparameter ranges considered for the machine learning models. The

† These decisions include ordering new ships, investing in second-hand ships, scrapping old and inefficient ships, operating or deactivating vessels (layup) depending on market conditions on the supply side, and hiring ships for transportation or not, and which route or commodity to trade on the demand side.

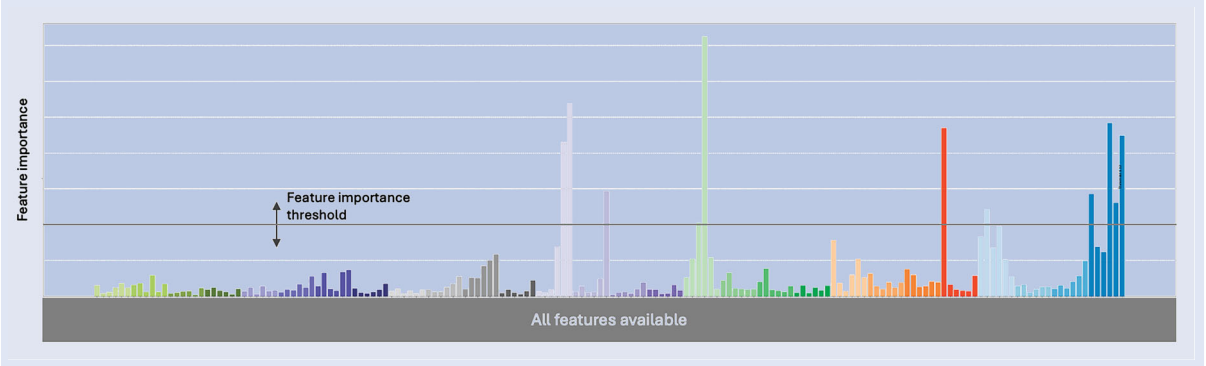


Figure 2. Feature importance and selection threshold.

Table 5. Final hyperparameters for CatBoost.

| Parameter | Value |
|---|---|
| Iterations[a] | 200 |
| Max depth[b] | 6 |
| Learning rate[c] | 0.1 |
| L2 regularization strength[d] | 1 |
| Random strength[e] | 1 |

[a]Number of trees.
[b]Maximum depth of each tree.
[c]Step size shrinkage.
[d]Strength of L2 norm regularization (log-scale).
[e]Randomness when choosing splits.

Table 6. Final hyperparameters for random forest.

| Parameter | Range |
|---|---|
| Estimators[a] | 50 |
| Max depth[b] | 5 |
| Min samples split[c] | 2 |
| Min samples leaf[d] | 1 |
| Max features[e] | 'log2' |

[a]Number of trees.
[b]'None' means max depth not constrained.
[c]Minimum number of samples required to split an internal node.
[d]Minimum number of samples required to be at a leaf node.
[e]The number of features to consider when looking for the best split. 'log2' is the base-2 logarithm of the total number of features.

Table 7. Final hyperparameters for XGBoost.

| Parameter | Range |
|---|---|
| Estimators[a] | 200 |
| Learning rate[b] | 0.01 |
| Max depth[c] | 10 |
| Subsample[d] | 0.8 |

[a]The number of boosting rounds/trees.
[b]Step size shrinkage.
[c]Maximum depth of each tree.
[d]Fraction of rows sampled for each boosting round

grid is iterated through by using the optimal set of features identified during the feature selection process, using the forward-validation scheme across five folds. The hyperparameter combination with the minimum weighted average MSE across the five folds is selected as the best set of hyperparameters, the same decision rule as for the feature selection. The pure time series machine learning configurations are inherently less computationally complex, as they use only the target value at time $t$ and potentially its lags up to a maximum of five as features. We determine the optimal lag structure (feature selection) and hyperparameters using the same forward-validation scheme across five folds described above. Final hyperparameters for CatBoost, Random Forest, and XGBoost are reported in tables 5, 6, and 7, respectively.

Model selection for the set of GARCH models is conducted using the Bayesian Information Criterion (BIC). All models are estimated by Maximum Likelihood methods, assuming either Gaussian or a Student-t error distribution in the first step. For each GARCH specification, the selection algorithm identifies the model with the lowest BIC among those which have successfully passed all post estimation diagnostic tests at 5% level of significance.

Table 8. Loss functions.

| Loss function | Formula | Type |
|---|---|---|
| $MSE_1$ | $T^{-1}\sum_{t=1}^{T}\left(\sigma_t-\hat{\sigma}_t\right)^2$ | Symmetric |
| $MSE_2$ | $T^{-1}\sum_{t=1}^{T}\left(\sigma_t^2-\hat{\sigma}_i^2\right)^2$ | Symmetric |
| QLIKE | $T^{-1}\sum_{t=1}^{T}\log\left(\hat{\sigma}_i^2\right)+y_i^2\hat{\sigma}_i^{-2}$ | Asymmetric |
| $R^2LOG$ | $T^{-1}\sum_{t=1}^{T}\left[\log\left(\sigma_t^2\hat{\sigma}_t^{-2}\right)\right]^2$ | Asymmetric |

Notes: $\hat{\sigma}_t$ denotes the predicted volatility for day t, $\sigma_t$ the conditional variance proxy, T the out-of-sample length.

## 5. The forecasting exercise

Volatility forecasting is particularly challenging when only daily data is available as volatility itself is latent and thus unobservable even ex-post. In general, to compare model based forecasts with ex-post realizations, the researcher must choose a statistical loss function and a proxy for the true unobservable conditional variance. When only noisy proxy

Table 9. SPA test results for the Baltic Handysize Average 7 Trip-charter rates (HA7TC).

| Benchmark | $MSE_1$ | | | $MSE_2$ | | | QLIKE | | | $R^2LOG$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ |
| 1-day-ahead | | | | | | | | | | | | |
| GARCH | 0.30 | **0.33** | 0.35 | 0.10 | **0.11** | 0.14 | 0.11 | **0.11** | 0.12 | 0.20 | **0.25** | 0.30 |
| EGARCH | 0.10 | **0.10** | 0.11 | 0.13 | **0.15** | 0.18 | 0.02 | 0.03 | 0.03 | 0.20 | **0.25** | 0.25 |
| GJR-GARCH | 0.15 | **0.15** | 0.15 | 0.30 | **0.33** | 0.33 | 0.11 | **0.12** | 0.13 | 0.13 | **0.15** | 0.16 |
| FIGARCH | 0.12 | **0.12** | 0.16 | 0.15 | **0.18** | 0.22 | 0.20 | **0.23** | 0.25 | 0.11 | **0.11** | 0.14 |
| GARCH-MIDAS | 0.01 | 0.05 | 0.05 | 0.04 | 0.05 | 0.06 | 0.04 | 0.06 | 0.11 | 0.03 | 0.04 | 0.05 |
| $RF_p$ | 0.06 | **0.10** | 0.13 | 0.11 | **0.11** | 0.11 | 0.12 | **0.12** | 0.20 | 0.30 | **0.33** | 0.40 |
| $RF_e$ | 0.01 | 0.02 | 0.03 | 0.01 | 0.02 | 0.05 | 0.04 | 0.06 | 0.06 | 0.04 | 0.04 | 0.06 |
| $XG_p$ | 0.02 | 0.02 | 0.00 | 0.01 | 0.04 | 0.03 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.05 |
| $XG_e$ | 0.05 | 0.06 | 0.08 | 0.01 | 0.02 | 0.03 | 0.04 | 0.04 | 0.06 | 0.04 | 0.05 | 0.05 |
| $CAT_p$ | 0.20 | **0.22** | 0.25 | 0.23 | **0.30** | 0.33 | 0.12 | **0.11** | 0.17 | 0.32 | **0.45** | 0.50 |
| $CAT_e$ | 0.11 | **0.11** | 0.18 | 0.04 | 0.05 | 0.05 | 0.01 | 0.02 | 0.02 | 0.01 | 0.02 | 0.03 |
| $SA_p$ | 0.18 | **0.25** | 0.28 | 0.19 | **0.30** | 0.30 | 0.18 | **0.28** | 0.30 | 0.30 | **0.33** | 0.35 |
| $SA_e$ | 0.35 | **0.35** | 0.37 | 0.19 | **0.23** | 0.26 | 0.18 | **0.27** | 0.30 | 0.42 | **0.56** | 0.63 |
| 5-days-ahead | | | | | | | | | | | | |
| GARCH | 0.43 | **0.65** | 0.69 | 0.46 | **0.77** | 0.77 | 0.27 | **0.27** | 0.27 | 0.63 | **0.84** | 0.89 |
| EGARCH | 0.60 | **0.89** | 0.91 | 0.53 | **0.84** | 0.85 | 0.19 | **0.33** | 0.33 | 1.00 | **1.00** | 1.00 |
| GJR-GARCH | 0.53 | **0.76** | 0.78 | 0.68 | **0.90** | 0.90 | 0.23 | **0.30** | 0.30 | 0.73 | **0.90** | 0.92 |
| FIGARCH | 0.10 | **0.12** | 0.12 | 0.11 | **0.11** | 0.11 | 0.23 | **0.25** | 0.25 | 0.24 | **0.27** | 0.33 |
| GARCH-MIDAS | 0.16 | **0.16** | 0.16 | 0.17 | **0.17** | 0.17 | 0.23 | **0.23** | 0.23 | 0.47 | **0.48** | 0.48 |
| $RF_p$ | 0.08 | 0.08 | 0.08 | 0.13 | **0.14** | 0.14 | 0.11 | **0.12** | 0.12 | 0.78 | **0.84** | 0.85 |
| $RF_e$ | 0.33 | **0.55** | 0.65 | 0.32 | **0.72** | 0.72 | 0.21 | **0.42** | 0.42 | 0.69 | **0.77** | 0.81 |
| $XG_p$ | 0.22 | **0.40** | 0.44 | 0.25 | **0.48** | 0.48 | 0.19 | **0.31** | 0.31 | 0.34 | **0.40** | 0.42 |
| $XG_e$ | 0.05 | 0.06 | 0.06 | 0.09 | 0.09 | 0.09 | 0.26 | **0.33** | 0.33 | 0.01 | 0.01 | 0.01 |
| $CAT_p$ | 0.27 | **0.61** | 0.63 | 0.44 | **0.84** | 0.86 | 0.15 | **0.30** | 0.30 | 0.70 | **0.78** | 0.80 |
| $CAT_e$ | 0.16 | **0.22** | 0.24 | 1.00 | **1.00** | 1.00 | 0.19 | **0.27** | 0.27 | 0.16 | **0.17** | 0.18 |
| $SA_p$ | 1.00 | **1.00** | 1.00 | 0.82 | **0.98** | 0.99 | 0.27 | **0.59** | 0.67 | 0.88 | **0.98** | 0.98 |
| $SA_e$ | 0.32 | **0.41** | 0.53 | 0.33 | **0.71** | 0.71 | 1.00 | **1.00** | 1.00 | 0.06 | 0.06 | 0.07 |
| 25-days-ahead | | | | | | | | | | | | |
| GARCH | 0.12 | **0.12** | 0.14 | 0.11 | **0.14** | 0.15 | 0.11 | **0.11** | 0.14 | 0.00 | 0.05 | 0.09 |
| EGARCH | 0.00 | 0.03 | 0.04 | 0.02 | 0.03 | 0.03 | 0.10 | **0.12** | 0.15 | 0.02 | 0.02 | 0.02 |
| GJR-GARCH | 0.12 | **0.15** | 0.16 | 0.15 | **0.18** | 0.19 | 0.12 | **0.12** | 0.14 | 0.06 | 0.08 | 0.08 |
| FIGARCH | 0.61 | **0.74** | 0.74 | 0.37 | **0.60** | 0.60 | 0.67 | **0.84** | 0.84 | 0.50 | **0.53** | 0.56 |
| GARCH-MIDAS | 0.42 | **0.42** | 0.42 | 0.45 | **0.45** | 0.45 | 0.23 | **0.23** | 0.26 | 0.83 | **0.85** | 0.85 |
| $RF_p$ | 0.02 | 0.03 | 0.03 | 0.01 | 0.01 | 0.02 | 0.03 | 0.03 | 0.03 | 0.01 | 0.01 | 0.03 |
| $RF_e$ | 0.61 | **0.87** | 0.89 | 0.31 | **0.62** | 0.64 | 0.45 | **0.50** | 0.54 | 0.36 | **0.43** | 0.48 |
| $XG_p$ | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 |
| $XG_e$ | 0.00 | 0.00 | 0.00 | 0.04 | 0.05 | 0.05 | 0.03 | 0.03 | 0.03 | 0.00 | 0.00 | 0.00 |
| $CAT_p$ | 0.15 | **0.15** | 0.20 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 |
| $CAT_e$ | 0.19 | **0.20** | 0.30 | 0.30 | **0.33** | 0.33 | 0.20 | **0.22** | 0.24 | 0.31 | **0.46** | 0.47 |
| $SA_p$ | 0.15 | **0.22** | 0.23 | 0.11 | **0.15** | 0.15 | 0.11 | **0.13** | 0.13 | 0.40 | **0.40** | 0.50 |
| $SA_e$ | 0.55 | **0.55** | 0.60 | 0.20 | **0.29** | 0.30 | 0.40 | **0.50** | 0.55 | 0.33 | **0.33** | 0.35 |

Note: RF - Random Forest; XG - XGBoost; CAT - CatBoost. Machine learning models denoted with subscript $p$ represent pure time series models, while models denoted with subscript $e$ incorporate exogenous variables as features. In each row, the benchmark model is assessed against the remaining set of models, where $p_C$ represents the consistent p-value, while $p_L$ and $p_U$ denote the lower and upper bounds, respectively. Bold consistent p-values indicate non-rejection of the null hypothesis at the 10% significance level.

are available such as daily data as in our case, as pointed out byAndersen *et al.* (2003) and Laurent *et al.* (2012), the use of a proxy might lead to a different ordering of competing models which would be obtained if the true volatility were observed. In our forecasting exercise, we follow Bauwens and Otranto (2016) and use several loss functions that are robust to noisy proxies. We define these in table 8. The MSE is a symmetric, quadratic loss function that is sensitive to outliers. Its use is advantageous if large errors should weight more heavily when assessing model performance. Unlike the symmetric MSE, the Quasi-Likelihood (QLIKE) loss function penalizes under-predictions more heavily than over-predictions. This makes QLIKE particularly useful in contexts where under-prediction is costly, i.e in risk management, as it favors positively biased forecasts. We evaluate the forecasting performance of the candidate models by means of two statistical approaches: the Superior Predictive Ability (SPA) test of Hansen (2005) and the Model Confidence Set (MCS) method of Hansen *et al.* (2011). The SPA test focuses on the predictive ability of a predefined benchmark model with respect to several alternatives: we employ it to assess if specific assumptions, such as the use of pure time series algorithms can be rejected, at the different horizons. As benchmarks, we choose the most parsimonious models taking into account the different assumptions in terms of exogenous predictors. This is aligned with industry practices where in

Table 10. SPA test results for the Baltic Supramax average 10 trip-charter rates (SA10TC).

| Benchmark | $MSE_1$ | | | $MSE_2$ | | | QLIKE | | | $R^2LOG$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ |
| | | | | | 1-day-ahead | | | | | | | |
| GARCH | 0.25 | **0.30** | 0.33 | 0.15 | **0.18** | 0.18 | 0.10 | **0.11** | 0.14 | 0.25 | **0.33** | 0.40 |
| EGARCH | 0.15 | **0.15** | 0.16 | 0.35 | **0.43** | 0.43 | 0.02 | 0.03 | 0.03 | 0.24 | **0.28** | 0.29 |
| GJR-GARCH | 0.12 | **0.15** | 0.15 | 0.34 | **0.37** | 0.37 | 0.10 | **0.11** | 0.11 | 0.28 | **0.48** | 0.48 |
| FIGARCH | 0.24 | **0.27** | 0.27 | 0.24 | **0.31** | 0.31 | 0.16 | **0.19** | 0.23 | 0.13 | **0.15** | 0.15 |
| GARCH-MIDAS | 0.08 | 0.08 | 0.08 | 0.09 | 0.09 | 0.09 | 0.04 | 0.06 | 0.11 | 0.03 | 0.05 | 0.08 |
| $RF_p$ | 0.06 | **0.10** | 0.13 | 0.11 | **0.11** | 0.11 | 0.10 | **0.15** | 0.18 | 0.30 | **0.52** | 0.50 |
| $RF_e$ | 0.01 | 0.03 | 0.06 | 0.03 | 0.03 | 0.05 | 0.08 | 0.08 | 0.09 | 0.04 | 0.04 | 0.06 |
| $XG_p$ | 0.02 | 0.02 | 0.02 | 0.0 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 | 0.04 | 0.05 | 0.05 |
| $XG_e$ | 0.07 | 0.07 | 0.07 | 0.02 | 0.02 | 0.05 | 0.06 | 0.07 | 0.08 | 0.06 | 0.07 | 0.07 |
| $CAT_p$ | 0.36 | **0.56** | 0.56 | 0.29 | **0.30** | 0.30 | 0.10 | **0.11** | 0.11 | 0.31 | **0.49** | 0.50 |
| $CAT_e$ | 0.11 | **0.11** | 0.18 | 0.04 | 0.05 | 0.05 | 0.01 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 |
| $SA_p$ | 0.18 | **0.21** | 0.28 | 0.19 | **0.25** | 0.30 | 0.11 | **0.21** | 0.21 | 0.31 | **0.33** | 0.33 |
| $SA_e$ | 0.35 | **0.35** | 0.37 | 0.19 | **0.23** | 0.26 | 0.18 | **0.27** | 0.30 | 0.42 | **0.56** | 0.63 |
| | | | | | 5-days-ahead | | | | | | | |
| GARCH | 0.43 | **0.65** | 0.69 | 0.46 | **0.77** | 0.77 | 0.27 | **0.27** | 0.27 | 0.63 | **0.84** | 0.89 |
| EGARCH | 0.60 | **0.89** | 0.91 | 0.53 | **0.84** | 0.85 | 0.19 | **0.33** | 0.33 | 1.00 | **1.00** | 1.00 |
| GJR-GARCH | 0.53 | **0.76** | 0.78 | 0.68 | **0.90** | 0.90 | 0.23 | **0.30** | 0.30 | 0.73 | **0.90** | 0.92 |
| FIGARCH | 0.10 | **0.12** | 0.12 | 0.11 | **0.11** | 0.11 | 0.23 | **0.25** | 0.25 | 0.24 | **0.27** | 0.33 |
| GARCH-MIDAS | 0.16 | **0.16** | 0.16 | 0.17 | **0.17** | 0.17 | 0.23 | **0.23** | 0.23 | 0.47 | **0.48** | 0.48 |
| $RF_p$ | 0.08 | 0.08 | 0.08 | 0.13 | **0.14** | 0.14 | 0.11 | **0.12** | 0.12 | 0.78 | **0.84** | 0.85 |
| $RF_e$ | 0.33 | **0.55** | 0.65 | 0.32 | **0.72** | 0.72 | 0.21 | **0.42** | 0.42 | 0.69 | **0.77** | 0.81 |
| $XG_p$ | 0.22 | **0.40** | 0.44 | 0.25 | **0.48** | 0.48 | 0.19 | **0.31** | 0.31 | 0.34 | **0.40** | 0.42 |
| $XG_e$ | 0.05 | 0.06 | 0.06 | 0.09 | 0.09 | 0.09 | 0.26 | **0.33** | 0.33 | 0.01 | 0.01 | 0.01 |
| $CAT_p$ | 0.27 | **0.61** | 0.63 | 0.44 | **0.84** | 0.86 | 0.15 | **0.30** | 0.30 | 0.70 | **0.78** | 0.80 |
| $CAT_e$ | 0.16 | **0.22** | 0.24 | 1.00 | **1.00** | 1.00 | 0.19 | **0.27** | 0.27 | 0.16 | **0.17** | 0.18 |
| $SA_p$ | 1.00 | **1.00** | 1.00 | 0.82 | **0.98** | 0.99 | 0.27 | **0.59** | 0.67 | 0.88 | **0.98** | 0.98 |
| $SA_e$ | 0.32 | **0.41** | 0.53 | 0.33 | **0.71** | 0.71 | 1.00 | **1.00** | 1.00 | 0.06 | 0.06 | 0.07 |
| | | | | | 25-days-ahead | | | | | | | |
| GARCH | 0.11 | **0.12** | 0.12 | 0.10 | **0.12** | 0.15 | 0.11 | **0.11** | 0.12 | 0.05 | 0.05 | 0.06 |
| EGARCH | 0.00 | 0.01 | 0.03 | 0.00 | 0.02 | 0.03 | 0.12 | **0.14** | 0.16 | 0.02 | 0.03 | 0.03 |
| GJR-GARCH | 0.14 | **0.16** | 0.17 | 0.11 | **0.12** | 0.15 | 0.11 | **0.12** | 0.15 | 0.06 | 0.08 | 0.08 |
| FIGARCH | 0.50 | **0.52** | 0.55 | 0.45 | **0.48** | 0.50 | 0.51 | **0.57** | 0.64 | 0.55 | **0.60** | 0.62 |
| GARCH-MIDAS | 0.50 | **0.55** | 0.65 | 0.38 | **0.42** | 0.45 | 0.30 | **0.33** | 0.33 | 0.71 | **0.71** | 0.74 |
| $RF_p$ | 0.00 | 0.01 | 0.03 | 0.01 | 0.02 | 0.05 | 0.02 | 0.04 | 0.05 | 0.01 | 0.01 | 0.03 |
| $RF_e$ | 0.61 | **0.87** | 0.89 | 0.31 | **0.62** | 0.64 | 0.45 | **0.50** | 0.54 | 0.36 | **0.43** | 0.48 |
| $XG_p$ | 0.00 | 0.01 | 0.01 | 0.00 | 0.01 | 0.02 | 0.01 | 0.02 | 0.04 | 0.01 | 0.01 | 0.05 |
| $XG_e$ | 0.00 | 0.01 | 0.01 | 0.02 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.00 | 0.02 | 0.04 |
| $CAT_p$ | 0.18 | **0.25** | 0.30 | 0.05 | 0.05 | 0.05 | 0.04 | 0.06 | 0.07 | 0.00 | 0.02 | 0.03 |
| $CAT_e$ | 0.12 | **0.12** | 0.15 | 0.10 | **0.12** | 0.15 | 0.20 | **0.25** | 0.30 | 0.30 | **0.33** | 0.37 |
| $SA_p$ | 0.10 | **0.12** | 0.16 | 0.11 | **0.12** | 0.14 | 0.10 | **0.15** | 0.18 | 0.37 | **0.40** | 0.44 |
| $SA_e$ | 0.67 | **0.70** | 0.71 | 0.30 | **0.33** | 0.38 | 0.42 | **0.50** | 0.58 | 0.51 | **0.65** | 0.60 |

Note: RF - Random Forest; XG - XGBoost; CAT - CatBoost. Machine learning models denoted with subscript $p$ represent pure time series models, while models denoted with subscript $e$ incorporate exogenous variables as features. In each row, the benchmark model is assessed against the remaining set of models, where $p_C$ represents the consistent p-value, while $p_L$ and $p_U$ denote the lower and upper bounds, respectively. Bold consistent p-values indicate non-rejection of the null hypothesis at the 10% significance level.

general very few predictors are used in freight rates forecasting. With the MCS method, we identify from the initial set of competing models those which display equal predictive ability and outperform the others at a given confidence level, offering guidance on the alternative models to use in practical forecasting industry applications.

The forecasting ability of the set of proposed models is evaluated over a series of 250 out-of-sample predictions.We compare the one-day, five-days and 25-days ahead volatility forecasts. Forecasts are constructed using a fixed rolling window scheme: the estimation period is rolled forward by adding one new daily observation and dropping the most distant observation. parameters are re-calibrated each day to obtain tomorrow's volatility forecasts and the sample size used for the estimation is fixed. Thus for each window, the input to the models includes the current set of training data plus the number of observations corresponding to the forecast step. This scheme satisfies the assumptions required by the MCS method of Hansen *et al.* (2011) and the SPA test of Hansen (2005) and allows a unified treatment of nested and unnested models, thus allowing us to compare machine learning and econometric models. For each statistical loss function, we evaluate the significance of the differences by means of the SPA test and the MCS methodology.

In the 'horse race', we consider several different models. We include: (i) all the GARCH models introduced in

Table 11. SPA test results for the Baltic Panamax average 5 trip-charter rates (PA5TC).

| Benchmark | $MSE_1$ | | | $MSE_2$ | | | QLIKE | | | $R^2LOG$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ |
| | | | | | 1-day-ahead | | | | | | | |
| GARCH | 0.33 | **0.54** | 0.57 | 0.10 | **0.11** | 0.15 | 0.15 | **0.14** | 0.16 | 0.25 | **0.31** | 0.33 |
| EGARCH | 0.14 | **0.16** | 0.16 | 0.33 | **0.42** | 0.46 | 0.03 | 0.03 | 0.04 | 0.27 | **0.30** | 0.30 |
| GJR-GARCH | 0.11 | **0.12** | 0.14 | 0.27 | **0.33** | 0.35 | 0.11 | **0.12** | 0.12 | 0.30 | **0.41** | 0.43 |
| FIGARCH | 0.22 | **0.28** | 0.29 | 0.20 | **0.26** | 0.30 | 0.15 | **0.15** | 0.18 | 0.15 | **0.17** | 0.20 |
| GARCH-MIDAS | 0.03 | 0.03 | 0.08 | 0.01 | 0.04 | 0.04 | 0.04 | 0.06 | 0.11 | 0.03 | 0.03 | 0.04 |
| $RF_p$ | 0.06 | **0.11** | 0.14 | 0.12 | **0.10** | 0.12 | 0.11 | **0.12** | 0.16 | 0.28 | **0.40** | 0.44 |
| $RF_e$ | 0.02 | 0.02 | 0.04 | 0.03 | 0.05 | 0.05 | 0.01 | 0.01 | 0.03 | 0.04 | 0.04 | 0.04 |
| $XG_p$ | 0.02 | 0.02 | 0.02 | 0.0 | 0.03 | 0.02 | 0.02 | 0.03 | 0.02 | 0.04 | 0.04 | 0.05 |
| $XG_e$ | 0.06 | 0.06 | 0.08 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.02 | 0.04 | 0.07 |
| $CAT_p$ | 0.18 | **0.25** | 0.26 | 0.27 | **0.27** | 0.30 | 0.10 | **0.11** | 0.12 | 0.30 | **0.33** | 0.36 |
| $CAT_e$ | 0.10 | **0.15** | 0.15 | 0.03 | 0.04 | 0.08 | 0.01 | 0.01 | 0.03 | 0.02 | 0.03 | 0.05 |
| $SA_p$ | 0.18 | **0.22** | 0.28 | 0.12 | **0.15** | 0.22 | 0.11 | **0.15** | 0.21 | 0.26 | **0.28** | 0.30 |
| $SA_e$ | 0.30 | **0.30** | 0.33 | 0.18 | **0.20** | 0.26 | 0.15 | **0.20** | 0.28 | 0.40 | **0.50** | 0.60 |
| | | | | | 5-days-ahead | | | | | | | |
| GARCH | 0.45 | **0.50** | 0.61 | 0.40 | **0.45** | 0.51 | 0.12 | **0.14** | 0.17 | 0.35 | **0.36** | 0.36 |
| EGARCH | 0.50 | **0.52** | 0.62 | 0.50 | **0.56** | 0.63 | 0.11 | **0.14** | 0.15 | 0.23 | **0.25** | 0.25 |
| GJR-GARCH | 0.43 | **0.50** | 0.56 | 0.35 | **0.35** | 0.40 | 0.33 | **0.33** | 0.35 | 0.34 | **0.35** | 0.35 |
| FIGARCH | 0.11 | **0.14** | 0.15 | 0.10 | **0.12** | 0.15 | 0.22 | **0.28** | 0.30 | 0.16 | **0.16** | 0.18 |
| GARCH-MIDAS | 0.16 | **0.18** | 0.18 | 0.10 | **0.11** | 0.15 | 0.12 | **0.13** | 0.15 | 0.30 | **0.33** | 0.33 |
| $RF_p$ | 0.08 | 0.08 | 0.09 | 0.11 | **0.11** | 0.14 | 0.10 | 0.11 **0.12** | 0.12 | 0.38 | **0.44** | 0.53 |
| $RF_e$ | 0.33 | **0.33** | 0.38 | 0.28 | **0.37** | 0.40 | 0.21 | **0.25** | 0.33 | 0.12 | **0.15** | 0.18 |
| $XG_p$ | 0.00 | 0.0.1 | 0.04 | 0.05 | 0.04 | 0.05 | 0.10 | **0.12** | 0.12 | 0.04 | 0.04 | 0.02 |
| $XG_e$ | 0.05 | 0.06 | 0.05 | 0.08 | 0.09 | 0.09 | 0.30 | **0.30** | 0.32 | 0.01 | 0.01 | 0.02 |
| $CAT_p$ | 0.27 | **0.61** | 0.63 | 0.44 | **0.84** | 0.86 | 0.15 | **0.30** | 0.30 | 0.70 | **0.78** | 0.80 |
| $CAT_e$ | 0.16 | **0.22** | 0.24 | 0.50 | **0.67** | 0.68 | 0.23 | **0.33** | 0.27 | 0.21 | **0.27** | 0.33 |
| $SA_p$ | 1.00 | **1.00** | 1.00 | 0.82 | **0.98** | 0.99 | 0.27 | **0.59** | 0.67 | 0.88 | **0.98** | 0.98 |
| $SA_e$ | 0.32 | **0.41** | 0.53 | 0.33 | **0.71** | 0.71 | 0.50 | **0.56** | 0.03 | 0.05 | 0.06 | 0.07 |
| | | | | | 25-days-ahead | | | | | | | |
| GARCH | 0.13 | **0.12** | 0.12 | 0.12 | **0.15** | 0.15 | 0.11 | **0.12** | 0.12 | 0.02 | 0.03 | 0.06 |
| EGARCH | 0.01 | 0.03 | 0.06 | 0.02 | 0.03 | 0.04 | 0.09 | **0.10** | 0.15 | 0.02 | 0.02 | 0.03 |
| GJR-GARCH | 0.12 | **0.12** | 0.12 | 0.15 | **0.15** | 0.18 | 0.10 | **0.15** | 0.16 | 0.04 | 0.02 | 0.03 |
| FIGARCH | 0.61 | **0.74** | 0.74 | 0.37 | **0.60** | 0.60 | 0.67 | **0.84** | 0.84 | 0.50 | **0.53** | 0.56 |
| GARCH-MIDAS | 0.42 | **0.42** | 0.42 | 0.45 | **0.45** | 0.45 | 0.23 | **0.23** | 0.26 | 0.83 | **0.85** | 0.85 |
| $RF_p$ | 0.02 | 0.03 | 0.03 | 0.01 | 0.01 | 0.02 | 0.03 | 0.03 | 0.03 | 0.01 | 0.01 | 0.03 |
| $RF_e$ | 0.61 | **0.87** | 0.89 | 0.31 | **0.62** | 0.64 | 0.45 | **0.50** | 0.54 | 0.36 | **0.43** | 0.48 |
| $XG_p$ | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 |
| $XG_e$ | 0.00 | 0.00 | 0.00 | 0.04 | 0.05 | 0.05 | 0.03 | 0.03 | 0.03 | 0.00 | 0.00 | 0.00 |
| $CAT_p$ | 0.14 | **0.17** | 0.17 | 0.03 | 0.03 | 0.04 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 |
| $CAT_e$ | 0.29 | **0.29** | 0.29 | 0.23 | **0.23** | 0.23 | 0.25 | **0.26** | 0.26 | 0.25 | **0.42** | 0.42 |
| $SA_p$ | 0.22 | **0.25** | 0.33 | 0.12 | **0.15** | 0.18 | 0.19 | **0.20** | 0.35 | 0.30 | **0.33** | 0.40 |
| $SA_e$ | 0.60 | **0.65** | 0.67 | 0.25 | **0.30** | 0.33 | 0.25 | **0.55** | 0.60 | 0.41 | **0.52** | 0.63 |

Note: RF - Random Forest; XG - XGBoost; CAT - CatBoost. Machine learning models denoted with subscript $p$ represent pure time series models, while models denoted with subscript $e$ incorporate exogenous variables as features. In each row, the benchmark model is assessed against the remaining set of models, where $p_C$ represents the consistent p-value, while $p_L$ and $p_U$ denote the lower and upper bounds, respectively. Bold consistent p-values indicate non-rejection of the null hypothesis at the 10% significance level.

Section 3 with Gaussian and t-distributed errors, (ii) the GARCH MIDAS models which uses the market tightness index as a macroeconomic volatility driver with Normal and t-distributed errors, (iii) and the machine learning models with exogenous features, (iv) the machine learning models with time series lags only. We also investigate whether combining the strengths of these models could be beneficial in term of forecasting accuracy. It is well known that combinations of forecasts from econometric models can achieve greater predictive accuracy (Timmermann 2013). Several combination strategies have been discussed in the literature, see X. Wang *et al.* (2023) for a review. In the machine learning literature, there is a growing consensus that ensembles (stacking) of learners (meta-learner) outperform methods that simply choose the best learner in term of predictive accuracy, due to their ability of capturing complex non-linear relationships between inputs and outputs (Reddy *et al.* 2022). Stacking involves using the predictions from base models as inputs for a meta-model (Aras 2021), and identifying the algorithm for learning at the meta-level (Džeroski and Ženko 2004).

In this article, we explore hybrid forecast combinations of GARCH models and tree-based learners. We tested the OLS regression, the LSTM and other learning algorithms, but these yield poor performance and low accuracy at the meta-level, due to the small size of the training set. To address this, we implement an equally weighted voting scheme that does not

involve additional learning at the meta-level (Džeroski and Ženko 2004). We consider several configurations of stacked models but present results in section 5 for the two best combinations. The first is the weighted average of all pure time series machine learning models included in the MCS at 75% level of significance for a given forecasting horizon and the best performing GARCH model at that horizon, the second is the weighted average of all machine learning algorithms with exogenous features included in the MCS at 75% level for a given forecasting horizon and the best performing GARCH model at that horizon. Thus the set of models included in the hybrid ensemble changes across the forecasting horizons. Results on the performance of other combinations and stacking strategies are available upon request from the authors. In the 'horse race' we include 110 models in total.

### *5.1. Assessing the benchmark: the SPA test*

In this section, we study the forecasting performance of a pre-specified benchmark model with respect to alternative models using the SPA test. The SPA test (Hansen 2005) directly compares a benchmark model to various alternative models according to a pre-specified loss function. The test enables us to shed light on the models characteristics which can be rejected. For a given loss function, it is based on the loss differential between the benchmark model, indexed by 0, and an alternative model $k = 1, \ldots, m$. The null hypothesis of the test is that the benchmark model is as good as any of the competitors in terms of expected loss. Every model is successively set as the benchmark and evaluated against the remaining set of models. Results for the different series are reported in tables 9 through 12 where $p_L$, $p_C$ and $p_U$ are, respectively, the consistent p-values and their lower and upper bounds. Consistent p-values are obtained by block bootstrap with 10 000 replications and a block length of $\sqrt{T}$, as suggested by Hansen (2005). Boldface entries indicate non-rejection of the null at the 10% significance level.

Across series, the results suggest that at the 1-day-ahead horizon, specifications including exogenous variables are rejected (GARCH-MIDAS, XGBoost and RF), while pure time series ones cannot be rejected, suggesting that at a short horizon the role of exogenous variables is not prominent. All the pure time series GARCH models are included for all the loss functions. Among the ML algorithms, the XGBoost is rejected across series. The stacked averages cannot be rejected for any series. At the 5 days ahead horizon, we reject again the XGBoost. For the GARCH models, the GARCH-MIDAS can no longer be rejected for the MSE loss functions. The CATBoost and the RF with exogenous variables are not rejected according to the MSE loss functions, while all ML specifications that don't include exogenous variables are rejected. All loss functions do not reject the stacked averages for any series. When considering the 25 days ahead horizon, the machine learning algorithms with no exogenous predictors and the stacked average with no exogenous predictors are all rejected. The RF, the CATBoost, the stacked average with exogenous predictors and the GARCH-MIDAS are included for all series for all loss functions, suggesting that the use of exogenous variables is very significant at longer horizon, regardless of the model structure used. GARCH models generally demonstrate more robust accuracy than their ML counterparts, with the GARCH-MIDAS specification being a particularly strong benchmark at the 25 days horizon. It is intriguing to note that the only model not rejected across all loss functions at any horizon is the stacked average with exogenous predictors.

### *5.2. The model confidence set*

The Model Confidence Set identifies a set of models with equivalent predictive ability that outperform all the other competing models at a given confidence level. The objective of the MCS procedure of Hansen *et al.* (2011) is to identify the optimal subset of models, $M^*$, from an initial set of competing models, $M^0$, at a predefined confidence level. $\hat{M}^* \subset M^0$ will encompass the models $M \in M^0$ that demonstrate the strongest relative forecasting performance according to a specific loss function. This method does not require prespecifying a preferred benchmark model; in fact, it is a statistical test of equivalence with respect to a particular loss function. The trimming is achieved via a sequence of equal predictive ability (EPA) tests. Hence, if the null hypothesis is rejected, the model with the poorest performance is removed from $M$. This sequential testing procedure continues until the null hypothesis of equal predictive ability is accepted at the given significance level and the Superior Set Models (MCS) $\hat{M}^*$ is obtained. For a large number of competing models, we follow Hansen *et al.* (2011) and obtain the quantiles of the asymptotic distribution of the test statistic by block bootstrap with 10 000 replications and a block length of $\sqrt{T}$. The results are illustrated in table 13, showing the inclusion rates of each model in the MCS at respectively 90% and 75% confidence levels at the various horizons.

The most striking result is the inclusion of the Stacked Average combination in the MCS of the four loss functions at all horizons, supporting the hypothesis that combinations of GARCH models and tree-based learners with exogenous variables improve forecasting accuracy at several horizons. This suggests that the weighting (voting) process helps mitigate the individual weaknesses and combine the strengths of different models. In line with our previous findings, machine learning models with exogenous features consistently exhibit good performance at medium and long horizon but do not outperform consistently the GARCH-MIDAS. This suggests that the market tightness index significantly enhances long-term forecast accuracy for these segments. For the BCI, the high inclusion rates for multiple models across various horizons indicate that a combination is more sensible. Overall, the MCS results suggest that machine learning models with exogenous features excel in long-term forecasting, but GARCH models are in general better at capturing the dependence structure of the volatility. Overall, our results suggest that a hybrid forecast combination enhances the predictive accuracy at all forecasting horizons.

### *5.3. Robustness*

In this section, we investigate the sensitivity of the models' forecasting performance with respect to the choice of

Table 12. SPA test results for the Baltic Capeszie average 5 trip-charter rates (CA5TC).

| Benchmark | $MSE_1$ | | | $MSE_2$ | | | QLIKE | | | $R^2LOG$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ | $p_L$ | $p_C$ | $p_U$ |
| | | | | | 1-day-ahead | | | | | | | |
| GARCH | 0.30 | **0.38** | 0.47 | 0.10 | **0.11** | 0.14 | 0.13 | **0.14** | 0.15 | 0.28 | **0.30** | 0.31 |
| EGARCH | 0.15 | **0.15** | 0.16 | 0.35 | **0.43** | 0.43 | 0.02 | 0.03 | 0.03 | 0.24 | **0.28** | 0.29 |
| GJR-GARCH | 0.12 | **0.15** | 0.15 | 0.34 | **0.37** | 0.37 | 0.10 | **0.11** | 0.11 | 0.28 | **0.48** | 0.48 |
| FIGARCH | 0.24 | **0.27** | 0.27 | 0.24 | **0.31** | 0.31 | 0.16 | **0.19** | 0.23 | 0.13 | **0.15** | 0.15 |
| GARCH-MIDAS | 0.08 | 0.08 | 0.08 | 0.09 | 0.09 | 0.09 | 0.04 | 0.06 | 0.11 | 0.03 | 0.05 | 0.08 |
| $RF_p$ | 0.06 | **0.10** | 0.13 | 0.11 | **0.11** | 0.11 | 0.10 | **0.15** | 0.18 | 0.30 | **0.52** | 0.50 |
| $RF_e$ | 0.01 | 0.03 | 0.06 | 0.03 | 0.03 | 0.05 | 0.08 | 0.08 | 0.09 | 0.04 | 0.04 | 0.06 |
| $XG_p$ | 0.02 | 0.02 | 0.02 | 0.0 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 | 0.04 | 0.05 | 0.05 |
| $XG_e$ | 0.07 | 0.07 | 0.07 | 0.02 | 0.02 | 0.05 | 0.06 | 0.07 | 0.08 | 0.06 | 0.07 | 0.07 |
| $CAT_p$ | 0.36 | **0.56** | 0.56 | 0.29 | **0.30** | 0.30 | 0.10 | **0.11** | 0.11 | 0.31 | **0.49** | 0.50 |
| $CAT_e$ | 0.11 | **0.11** | 0.18 | 0.04 | 0.05 | 0.05 | 0.01 | 0.02 | 0.02 | 0.02 | 0.03 | 0.03 |
| $SA_p$ | 0.18 | **0.21** | 0.28 | 0.19 | **0.25** | 0.30 | 0.11 | **0.21** | 0.21 | 0.31 | **0.33** | 0.33 |
| $SA_e$ | 0.35 | **0.35** | 0.37 | 0.19 | **0.23** | 0.26 | 0.18 | **0.27** | 0.30 | 0.42 | **0.56** | 0.63 |
| | | | | | 5-days-ahead | | | | | | | |
| GARCH | 0.43 | **0.65** | 0.69 | 0.46 | **0.77** | 0.77 | 0.27 | **0.27** | 0.27 | 0.63 | **0.84** | 0.89 |
| EGARCH | 0.60 | **0.89** | 0.91 | 0.53 | **0.84** | 0.85 | 0.19 | **0.33** | 0.33 | 1.00 | **1.00** | 1.00 |
| GJR-GARCH | 0.53 | **0.76** | 0.78 | 0.68 | **0.90** | 0.90 | 0.23 | **0.30** | 0.30 | 0.73 | **0.90** | 0.92 |
| FIGARCH | 0.10 | **0.12** | 0.12 | 0.11 | **0.11** | 0.11 | 0.23 | **0.25** | 0.25 | 0.24 | **0.27** | 0.33 |
| GARCH-MIDAS | 0.16 | **0.16** | 0.16 | 0.17 | **0.17** | 0.17 | 0.23 | **0.23** | 0.23 | 0.47 | **0.48** | 0.48 |
| $RF_p$ | 0.08 | 0.08 | 0.08 | 0.13 | **0.14** | 0.14 | 0.11 | **0.12** | 0.12 | 0.78 | **0.84** | 0.85 |
| $RF_e$ | 0.33 | **0.55** | 0.65 | 0.32 | **0.72** | 0.72 | 0.21 | **0.42** | 0.42 | 0.69 | **0.77** | 0.81 |
| $XG_p$ | 0.22 | **0.40** | 0.44 | 0.25 | **0.48** | 0.48 | 0.19 | **0.31** | 0.31 | 0.34 | **0.40** | 0.42 |
| $XG_e$ | 0.05 | 0.06 | 0.06 | 0.09 | 0.09 | 0.09 | 0.26 | **0.33** | 0.33 | 0.01 | 0.01 | 0.01 |
| $CAT_p$ | 0.27 | **0.61** | 0.63 | 0.44 | **0.84** | 0.86 | 0.15 | **0.30** | 0.30 | 0.70 | **0.78** | 0.80 |
| $CAT_e$ | 0.16 | **0.22** | 0.24 | 1.00 | **1.00** | 1.00 | 0.19 | **0.27** | 0.27 | 0.16 | **0.17** | 0.18 |
| $SA_p$ | 1.00 | **1.00** | 1.00 | 0.82 | **0.98** | 0.99 | 0.27 | **0.59** | 0.67 | 0.88 | **0.98** | 0.98 |
| $SA_e$ | 0.32 | **0.41** | 0.53 | 0.33 | **0.71** | 0.71 | 1.00 | **1.00** | 1.00 | 0.06 | 0.06 | 0.07 |
| | | | | | 25-days-ahead | | | | | | | |
| GARCH | 0.13 | **0.12** | 0.12 | 0.12 | **0.15** | 0.15 | 0.11 | **0.12** | 0.12 | 0.02 | 0.03 | 0.06 |
| EGARCH | 0.01 | 0.03 | 0.06 | 0.02 | 0.03 | 0.04 | 0.09 | **0.10** | 0.15 | 0.02 | 0.02 | 0.03 |
| GJR-GARCH | 0.12 | **0.12** | 0.12 | 0.15 | **0.15** | 0.18 | 0.10 | **0.15** | 0.16 | 0.04 | 0.02 | 0.03 |
| FIGARCH | 0.61 | **0.74** | 0.74 | 0.37 | **0.60** | 0.60 | 0.67 | **0.84** | 0.84 | 0.50 | **0.53** | 0.56 |
| GARCH-MIDAS | 0.42 | **0.42** | 0.42 | 0.45 | **0.45** | 0.45 | 0.23 | **0.23** | 0.26 | 0.83 | **0.85** | 0.85 |
| $RF_p$ | 0.02 | 0.03 | 0.03 | 0.01 | 0.01 | 0.02 | 0.03 | 0.03 | 0.03 | 0.01 | 0.01 | 0.03 |
| $RF_e$ | 0.61 | **0.87** | 0.89 | 0.31 | **0.62** | 0.64 | 0.45 | **0.50** | 0.54 | 0.36 | **0.43** | 0.48 |
| $XG_p$ | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 | 0.02 | 0.01 | 0.01 | 0.02 |
| $XG_e$ | 0.00 | 0.00 | 0.00 | 0.04 | 0.05 | 0.05 | 0.03 | 0.03 | 0.03 | 0.00 | 0.00 | 0.00 |
| $CAT_p$ | 0.14 | **0.17** | 0.17 | 0.03 | 0.03 | 0.04 | 0.01 | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 |
| $CAT_e$ | 0.29 | **0.29** | 0.29 | 0.23 | **0.23** | 0.23 | 0.25 | **0.26** | 0.26 | 0.25 | **0.42** | 0.42 |
| $SA_p$ | 0.15 | **0.19** | 0.23 | 0.11 | **0.11** | 0.15 | 0.11 | **0.11** | 0.14 | 0.41 | **0.47** | 0.54 |
| $SA_e$ | 0.64 | **0.84** | 0.88 | 0.24 | **0.49** | 0.59 | 0.26 | **0.56** | 0.66 | 0.49 | **0.62** | 0.73 |

Note: RF - Random Forest; XG - XGBoost; CAT - CatBoost. Machine learning models denoted with subscript $p$ represent pure time series models, while models denoted with subscript $e$ incorporate exogenous variables as features. In each row, the benchmark model is assessed against the remaining set of models, where $p_C$ represents the consistent p-value, while $p_L$ and $p_U$ denote the lower and upper bounds, respectively. Bold consistent p-values indicate non-rejection of the null hypothesis at the 10% significance level.

the forecasting sample. The time span of our sample covers the period from 2017 to 2023 and several different volatility dynamics. Hansen *et al.* (2003) point out that the SPA test and MCS approach are specific to the set of candidate models and the sample period. Here, we investigate the sensitivity of the models' forecasting performance with respect to the forecast evaluation sample based on two sub-samples which are homogeneous in their volatility dynamics. The choice of periods reflects the dynamics of freight rates and shipping markets. The first sub-sample, from the beginning of our sample till July 2019, corresponds to a relatively calm period for the market as opposed to later periods. Our second sub-sample, from January 2020 to 2023, arguably represents the most turbulent period in shipping market history. This period encompasses several global events marked by unprecedented market uncertainty, including the pandemic and its aftermath, as well as the war in Ukraine. The volatility dynamics are very different in these two periods.

As expected, there are differences from the full sample results, but our findings support the forecasting benefit of combinations. The results are illustrated in tables 14 and 15. In a period of relatively calm markets, the combination method does not over-perform the GARCH-MIDAS model according to all loss functions. These periods are characterized by relatively small and slow-moving changes in volatility, as well as few changes in the macroeconomic drivers that

Table 13. Percentage inclusion in the MCS at the 90% and 75% level.

| Series | HA7TC | | | SA10TC | | | PA5TC | | | CA5TC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Forecast step | 1 | 5 | 25 | 1 | 5 | 25 | 1 | 5 | 25 | 1 | 5 | 25 |
| Confidence level: 90% | | | | | | | | | | | | |
| **Model** | | | | | | | | | | | | |
| GARCH | 80 | 83 | 18 | 75 | 84 | 33 | 33 | 33 | 17 | 58 | 58 | 50 |
| EGARCH | 20 | 18 | 10 | 25 | 15 | | 15 | 20 | 20 | 15 | 10 | 10 |
| GJR-GARCH | 55 | 53 | 45 | 37 | 40 | 10 | 33 | 33 | 20 | 89 | 88 | 33 |
| FIGARCH | 41 | 33 | 30 | 50 | 44 | 30 | 50 | 43 | 33 | 95 | 73 | 28 |
| GARCH-MIDAS | 10 | 33 | 58 | 17 | 34 | 48 | 15 | 33 | 58 | 5 | 33 | 83 |
| $RF_p$ | 35 | 17 | 5 | 50 | 1 | 0 | 22 | 17 | 10 | 80 | 80 | 5 |
| $RF_e$ | 20 | 30 | 67 | 23 | 33 | 52 | 17 | 22 | 58 | 15 | 33 | 45 |
| $XG_p$ | 0 | 0 | 0 | 5 | 8 | 0 | 0 | 0 | 0 | 9 | 10 | 1 |
| $XG_e$ | 5 | 3 | 7 | 3 | 13 | 15 | 1 | 7 | 0 | 1 | 3 | 10 |
| $CAT_p$ | 33 | 33 | 0 | 20 | 18 | 10 | 0 | 34 | 30 | 22 | 18 | 10 |
| $CAT_e$ | 45 | 53 | 58 | 5 | 37 | 53 | 14 | 34 | 63 | 12 | 18 | 65 |
| $SA_p$ | 80 | 80 | 60 | 50 | 83 | 35 | 78 | 65 | 43 | 65 | 70 | 33 |
| $SA_e$ | 15 | 50 | 85 | 18 | 83 | 85 | 20 | 83 | 85 | 15 | 68 | 90 |
| Confidence level: 75% | | | | | | | | | | | | |
| **Model** | | | | | | | | | | | | |
| GARCH | 89 | 83 | 67 | 85 | 83 | 52 | 78 | 21 | 67 | 78 | 53 | 10 |
| EGARCH | 10 | 13 | 3 | 5 | 5 | 0 | 3 | 3 | 2 | 5 | 0 | |
| GJR-GARCH | 81 | 83 | 67 | 85 | 80 | 67 | 85 | 83 | 67 | 78 | 55 | 58 |
| FIGARCH | 33 | 33 | 78 | 33 | 50 | 83 | 50 | 83 | 83 | 55 | 81 | 82 |
| GARCH-MIDAS | 10 | 33 | 87 | 5 | 33 | 84 | 2 | 3 | 78 | 11 | 37 | 83 |
| $RF_p$ | 47 | 30 | 20 | 68 | 33 | 16 | 43 | 30 | 15 | 55 | 42 | 17 |
| $RF_e$ | 17 | 37 | 40 | 13 | 20 | 37 | 18 | 25 | 39 | 15 | 25 | 36 |
| $XG_p$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 0 |
| $XG_e$ | 33 | 17 | 100 | 33 | 0 | 33 | 97 | 29 | 0 | 55 | 55 | 0 |
| $CAT_p$ | 67 | 17 | 15 | 43 | 16 | 0 | 47 | 23 | 10 | 72 | 33 | 10 |
| $CAT_e$ | 57 | 63 | 77 | 33 | 34 | 83 | 29 | 47 | 83 | 28 | 60 | 76 |
| $SA_p$ | 83 | 50 | 33 | 83 | 50 | 16 | 65 | 23 | 17 | 58 | 33 | 10 |
| $SA_e$ | 33 | 50 | 88 | 28 | 50 | 88 | 30 | 67 | 83 | 33 | 67 | 98 |

Note: RF - Random Forest; XG - XGBoost; CAT - CatBoost. Machine learning models denoted with subscript *p* represent pure time series models, while models denoted with subscript *e* incorporate exogenous variables as features. The percentages represent the inclusion rate of each model in the MCS across the loss functions.

impact the volatility dynamics. However, in periods of high turbulence, using the forecasting combination improves performance more than in the full sample.

## 6. Conclusion

Accurate volatility forecasts are key inputs in risk management processes in any market. The shipping freight market is and extremely volatile sector of the economy and accurate predictions are crucial for risk management capabilities of shipping market participants. Empirical evidence suggests that freight rates volatility may be driven by certain industry and macroeconomic variables.

The current literature has extensively investigated the use GARCH models to predict the volatility dynamics of freight rates. Machine learning algorithms have been successfully applied to forecast volatility in other asset classes. GARCH models can include only few additional covariates in volatility prediction, thus it is sensible to investigate whether ML approaches with their inherent capabilities to deal with high-dimensional predictors with complex inter-dependencies can improve prediction accuracy. In this paper, we conduct an extensive forecasting comparison of GARCH-type models including the GARCH-MIDAS, and tree-based machine learning methods ( Random Forest, XGBoost and CatBoost) to investigate the benefits of different approaches at different forecasting horizons. In our comprehensive analysis of the out-of-sample performance, we also test hybrid horizon-varying forecast combinations to investigate whether combining the strengths of ML and GARCH models could be beneficial in terms of forecasting accuracy. We use an equally weighted voting scheme to obtain the Stacked Average combination, combining the best models in the Model Confidence Set at 75% level for each horizon.

We consider and examine the effect of several exogenous variables which reflect the supply and the demand-side of the freight market as predictors of volatility. Moreover, we define a custom index, the Market Tightness Index, that proxies market conditions through sector-specific supply and demand macroeconomic features including seaborne trade, fleet size, and fuel oil prices. For the machine learning algorithms, we incorporate low-frequency exogenous features by Denton-Cholette transformation—an approach that, to our knowledge, has not previously been applied in forecasting freight market volatility.

Table 14. Nov. 2017–Jul. 2019: Percentage inclusion in the MCS at the 90% and 75% level.

| Series | HA7TC | | | SA10TC | | | PA5TC | | | CA5TC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Forecast step | 1 | 5 | 25 | 1 | 5 | 25 | 1 | 5 | 25 | 1 | 5 | 25 |
| Confidence level: 90% | | | | | | | | | | | | |
| **Model** | | | | | | | | | | | | |
| GARCH | 80 | 83 | 18 | 75 | 84 | 33 | 33 | 33 | 17 | 58 | 58 | 50 |
| EGARCH | 20 | 18 | 10 | 25 | 15 | | 15 | 20 | 20 | 15 | 10 | 10 |
| GJR-GARCH | 55 | 53 | 45 | 37 | 40 | 10 | 33 | 33 | 20 | 89 | 88 | 33 |
| FIGARCH | 41 | 33 | 30 | 50 | 44 | 30 | 50 | 43 | 33 | 95 | 73 | 28 |
| GARCH-MIDAS | 10 | 33 | 58 | 17 | 34 | 48 | 15 | 33 | 58 | 5 | 33 | 83 |
| $RF_p$ | 35 | 17 | 5 | 50 | 1 | 0 | 22 | 17 | 10 | 80 | 80 | 5 |
| $RF_e$ | 20 | 30 | 67 | 23 | 33 | 52 | 17 | 22 | 58 | 15 | 33 | 45 |
| $XG_p$ | 0 | 0 | 0 | 5 | 8 | 0 | 0 | 0 | 0 | 9 | 10 | 1 |
| $XG_e$ | 5 | 3 | 7 | 3 | 13 | 15 | 1 | 7 | 0 | 1 | 3 | 10 |
| $CAT_p$ | 33 | 33 | 0 | 20 | 18 | 10 | 0 | 34 | 30 | 22 | 18 | 10 |
| $CAT_e$ | 45 | 53 | 58 | 5 | 37 | 53 | 14 | 34 | 63 | 12 | 18 | 65 |
| $SA_p$ | 80 | 80 | 60 | 50 | 83 | 35 | 78 | 65 | 43 | 65 | 70 | 33 |
| $SA_e$ | 15 | 50 | 85 | 18 | 83 | 85 | 20 | 83 | 85 | 15 | 68 | 90 |
| Confidence level: 75% | | | | | | | | | | | | |
| **Model** | | | | | | | | | | | | |
| GARCH | 89 | 83 | 67 | 85 | 83 | 52 | 78 | 21 | 67 | 78 | 53 | 10 |
| EGARCH | 10 | 13 | 3 | 5 | 5 | 0 | 3 | 3 | 2 | 5 | 0 | |
| GJR-GARCH | 81 | 83 | 67 | 85 | 80 | 67 | 85 | 83 | 67 | 78 | 55 | 58 |
| FIGARCH | 33 | 33 | 78 | 33 | 50 | 83 | 50 | 83 | 83 | 55 | 81 | 82 |
| GARCH-MIDAS | 87 | 90 | 87 | 90 | 65 | 84 | 2 | 3 | 78 | 65 | 80 | 83 |
| $RF_p$ | 47 | 30 | 20 | 68 | 33 | 16 | 43 | 30 | 15 | 55 | 42 | 17 |
| $RF_e$ | 17 | 37 | 40 | 13 | 20 | 37 | 18 | 25 | 39 | 15 | 25 | 36 |
| $XG_p$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 3 | 0 |
| $XG_e$ | 33 | 17 | 100 | 33 | 0 | 0 | 15 | 29 | 0 | 55 | 55 | 0 |
| $CAT_p$ | 67 | 17 | 0 | 43 | 16 | 0 | 47 | 23 | 10 | 72 | 33 | 10 |
| $CAT_e$ | 57 | 63 | 77 | 33 | 34 | 83 | 29 | 47 | 83 | 28 | 60 | 50 |
| $SA_p$ | 83 | 50 | 65 | 83 | 50 | 55 | 65 | 23 | 17 | 58 | 33 | 10 |
| $SA_e$ | 33 | 50 | 88 | 50 | 50 | 88 | 55 | 67 | 83 | 33 | 67 | 98 |

Note: Results on the relatively calm sub-sample. RF - Random Forest; XG - XGBoost; CAT - CatBoost. Machine learning models denoted with subscript *p* represent pure time series models, while models denoted with subscript *e* incorporate exogenous variables as features. The percentages represent the inclusion rate of each model in the MCS across the loss functions.

The forecasting ability of the set of proposed models is evaluated over a series of 250 out-of-sample predictions for volatility over 1-day, 5-days and 25-days ahead, based on a fixed rolling window scheme. This scheme satisfies the assumptions required by the MCS method of Hansen *et al.* (2011) and the SPA test of Hansen (2005), which allows a unified treatment of nested and unnested models, and is appropriate for comparing the performance of machine learning and econometric models. Volatility forecasts are evaluated by means of the SPA test and the MCS methodology using symmetric and asymmetric loss functions, robust to the choice of the volatility proxy.

The results indicate that machine learning models with exogenous features consistently perform well at medium and long horizons but do not outperform consistently at these horizons the GARCH-MIDAS in predicting volatility of freight rates. The most striking result is the inclusion of the Stacked Average combination in the MCS of all the loss functions at all horizons, supporting the hypothesis that combinations of GARCH models and tree-based learners (CATboost and RF) with exogenous variables improve forecasting accuracy at several horizons. This suggests that the weighting (voting) process helps mitigate individual weaknesses and combine the strengths of different models. The findings could be attributed to the slower reaction of the shipping freight market to macro factors (exogenous variables) in the short-term, overset by their significant impact over longer periods. The Stacked Average combination jointly exploits the tree-based algorithm's ability to extract substantial incremental information about future volatility from predictors and the GARCH model ability to capture, in a parsimonious and effective way, the time dependence of the volatility.

The reported results have important implications for risk modeling and assessment in dry bulk shipping, as well as risk management practices of agents, including shipowners, charterers, and traders. For instance, more accurate forecast of freight rate volatility can help shipowners and charterers better assess their risk exposure and Value-at-Risk estimates, enhance their operational portfolios, improve budget planning, and manage costs. In addition, accurate assessment of freight rate volatility is necessary for pricing freight derivatives and related risk management instruments. Finally, a better and more accurate forecast of freight rate volatility can improve the estimation of hedge ratios and the implementation of hedging strategies to manage freight rate risk efficiently and effectively.

Table 15. Jan. 2020 to Sep. 2023: Percentage inclusion in the MCS at the 90% and 75% level.

| Series | HA7TC | | | SA10TC | | | PA5TC | | | CA5TC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Forecast step | 1 | 5 | 25 | 1 | 5 | 25 | 1 | 5 | 25 | 1 | 5 | 25 |
| Confidence level: 90% | | | | | | | | | | | | |
| **Model** | | | | | | | | | | | | |
| GARCH | 80 | 83 | 18 | 75 | 84 | 33 | 33 | 33 | 17 | 58 | 58 | 50 |
| EGARCH | 20 | 18 | 10 | 25 | 15 | | 15 | 20 | 20 | 15 | 10 | 10 |
| GJR-GARCH | 55 | 53 | 45 | 37 | 40 | 10 | 33 | 33 | 20 | 89 | 88 | 33 |
| FIGARCH | 41 | 33 | 30 | 50 | 44 | 30 | 50 | 43 | 33 | 95 | 73 | 28 |
| GARCH-MIDAS | 10 | 33 | 58 | 17 | 34 | 48 | 15 | 33 | 58 | 5 | 33 | 83 |
| $RF_p$ | 35 | 17 | 5 | 50 | 1 | 0 | 22 | 17 | 10 | 80 | 80 | 5 |
| $RF_e$ | 20 | 30 | 67 | 23 | 33 | 52 | 17 | 22 | 58 | 15 | 33 | 45 |
| $XG_p$ | 0 | 0 | 0 | 5 | 8 | 0 | 0 | 0 | 0 | 9 | 10 | 1 |
| $XG_e$ | 5 | 3 | 7 | 3 | 13 | 15 | 1 | 7 | 0 | 1 | 3 | 10 |
| $CAT_p$ | 33 | 33 | 0 | 20 | 18 | 10 | 0 | 34 | 30 | 22 | 18 | 10 |
| $CAT_e$ | 45 | 53 | 58 | 5 | 37 | 53 | 14 | 34 | 63 | 12 | 18 | 65 |
| $SA_p$ | 80 | 80 | 60 | 50 | 83 | 35 | 78 | 65 | 43 | 65 | 70 | 33 |
| $SA_e$ | 15 | 50 | 85 | 18 | 83 | 85 | 20 | 83 | 85 | 15 | 68 | 90 |
| Confidence level: 75% | | | | | | | | | | | | |
| **Model** | | | | | | | | | | | | |
| GARCH | 67 | 56 | 67 | 49 | 83 | 52 | 78 | 21 | 67 | 65 | 53 | 33 |
| EGARCH | 0 | 13 | 3 | 5 | 15 | 17 | 3 | 3 | 2 | 5 | 0 | |
| GJR-GARCH | 81 | 83 | 67 | 85 | 76 | 67 | 85 | 83 | 67 | 78 | 33 | 33 |
| FIGARCH | 33 | 33 | 33 | 33 | 50 | 50 | 50 | 50 | 33 | 55 | 50 | 33 |
| GARCH-MIDAS | 10 | 33 | 87 | 5 | 33 | 84 | 2 | 3 | 78 | 11 | 37 | 83 |
| $RF_p$ | 47 | 30 | 20 | 35 | 33 | 16 | 43 | 30 | 15 | 55 | 42 | 26 |
| $RF_e$ | 17 | 37 | 40 | 13 | 20 | 37 | 18 | 25 | 39 | 15 | 25 | 33 |
| $XG_p$ | 0 | 0 | 0 | 33 | 0 | 16 | 0 | 0 | 0 | 1 | 3 | 0 |
| $XG_e$ | 33 | 17 | 100 | 33 | 0 | 33 | 97 | 29 | 0 | 55 | 55 | 0 |
| $CAT_p$ | 67 | 17 | 15 | 43 | 16 | 0 | 47 | 23 | 10 | 72 | 33 | 10 |
| $CAT_e$ | 57 | 63 | 32 | 33 | 34 | 83 | 29 | 47 | 33 | 28 | 60 | 21 |
| $SA_p$ | 88 | 76 | 98 | 86 | 50 | 67 | 65 | 23 | 75 | 58 | 63 | 78 |
| $SA_e$ | 76 | 75 | 88 | 78 | 75 | 88 | 76 | 67 | 83 | 59 | 67 | 98 |

Note: The tables refer to high volatility periods. RF - Random Forest; XG - XGBoost; CAT - CatBoost. Machine learning models denoted with subscript *p* represent pure time series models, while models denoted with subscript *e* incorporate exogenous variables as features. The percentages represent the inclusion rate of each model in the MCS across the loss functions.

It is worth noting that our findings may significantly change if intra-day data on freight rates were to become available. The availability of high-frequency data would allow researchers to explore the forecasting performances of realized volatility and deep learning models, such as LSTM, which cannot be used with daily frequency data only.

## Acknowledgments

We are grateful for helpful comments from participants at the Lillehammer Business Analytics Conference, 22–24 May 2024. Furthermore, we thank the Editor and two anonymous reviewers for insightful comments that helped us further improve the paper.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Morten Risstad acknowledges partial funding of his contribution to this research by The Research Council of Norway throughout the project COMPAMA (https://www.ntnu.edu/compama/), with grant number 314609.

## References

Abouarghoub , W., Mariscal, I.B. and Howells, P., A two-state Markov-switching distinctive conditional variance application for tanker freight returns. *Int. J. Financ. Eng. Risk Manag.*, 2014, **1**, 239–263.

Agnolucci, P., Volatility in crude oil futures: A comparison of the predictive ability of GARCH and implied volatility models. *Energy Econ.*, 2009, **31**, 316–321.

Alizadeh, A.H. and Nomikos, N.K., *Shipping Derivatives and Risk Management*, 2009 (London: Palgrave Macmilan).

Alizadeh, A. and Nomikos, N., Dynamics of the term structure and volatility of shipping freight rates. *J. Transp. Econ. Policy*, 2011, **45**, 105–128.

Alizadeh, A.H. and Sun, X, Hedging shipping freight rates using conditional value-at-risk and buffered probability of exceedance. Working Paper, 2023.

Alizadeh, A.H., Huang, C.Y. and van Dellen, S., A regime switching approach for hedging tanker shipping freight rates. *Energy Econ.*, 2015, **49**, 44–59.

Andersen, T. and Bollerslev, T., Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *Int. Econ. Rev.*, 1998, **39**(4), 885–905.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Labys, P., Modeling and forecasting realized volatility. *Econometrica*, 2003, **71**, 579–625.

Anderson, T.G., Bollerslev, T., Christoffersen, P.F. and Diebold, F.X, Volatility and correlation forecasting. In *Handbook of Economic Forecasting, Volume 1*, 2006.

Angelidis, T. and Skiadopoulos, G., Measuring the market risk of freight rates: A value-at-risk approach. *Int. J. Theor. Appl. Finance*, 2008, **11**, 447–469.

Aras, S., Stacking hybrid GARCH models for forecasting Bitcoin volatility. *Expert Syst. Appl.*, 2021, **174**, 114747.

Argyropoulos, C. and Panopoulou, E., Measuring the market risk of freight rates: A forecast combination approach. *J. Forecast.*, 2018, **37**, 201–224.

Baltic Exchange, Dry, 2023.

Bauwens, L. and Otranto, E., Modeling the dependence of conditional correlations on market volatility. *J. Bus. Econ. Stat.*, 2016, **34**, 254–268.

Bentes, S.R., Forecasting volatility in gold returns under the GARCH, IGARCH and FIGARCH frameworks: New evidence. *Physica A*, 2015, **438**, 355–364.

Bergmeir, C., Hyndman, R.J. and Koo, B., A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Comput. Stat. Data Anal.*, 2018, **120**, 70–83.

Boudt, K., Danielsson, J. and Laurent, S., Robust forecasting of dynamic conditional correlation GARCH models. *Int. J. Forecast.*, 2013, **29**, 244–257.

Brownlee, J, *XGBoost with Python - Gradient Boosted Trees with XGBoost and scikit-learn*, 2017 (Machine Learning Mastery).

Che, Z., Purushotham, S., Cho, K., Sontag, D. and Liu, Y., Recurrent neural networks for multivariate time series with missing values. *Sci. Rep.*, 2018, **8**, 6085.

Chen, G. and Chen, J., A novel wrapper method for feature selection and its applications. *Neurocomputing*, 2015, **159**, 219–226.

Chen, T. and Guestrin, C, Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016 (Association for Computing Machinery).

Christensen, K., Siggaard, M. and Veliyev, B., A machine learning approach to volatility forecasting. *J. Financ. Econom.*, 2023, **21**, 1680–1727.

Drobetz, W., Richter, T. and Wambach, M., Dynamics of time-varying volatility in the dry bulk and tanker freight markets. *Appl. Financ. Econ.*, 2012, **22**, 1367–1384.

Džeroski, S. and Ženko, B., Is combining classifiers with stacking better than selecting the best one? *Mach. Learn.*, 2004, **54**, 255–273.

Engle, R.F., Autoregressive conditional heteroskedasticity with estimates of the variance of U.K. inflation. *Econometrica*, 1982, **50**, 987–1008.

Engle, R.F. and Lee, G, A permanent and transitory component model of stock return volatility. In *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive*, edited by W.J. Granger, R.F. Engle and H. White, 1999 (Oxford University Press: Oxford, UK).

Engle, R.F., Ghysels, E. and Sohn, B., Stock market volatility and macroeconomic fundamentals. *Rev. Econ. Stat.*, 2013, **95**, 776–797.

Franq, C. and Zakoian, J.M., *GARCH Models: Structure, Statistical Inference and Financial Applications*, 2019 (Wiley).

Franses, P.H. and Van Dijk, D., Forecasting stock market volatility using (non-linear) Garch models. *J. Forecast.*, 1996, **15**, 229–235.

Gavriilidis, K., Kambouroudis, D.S., Tsakou, K. and Tsouknidis, D.A., Volatility forecasting across tanker freight rates: The role of oil price shocks. *Transp. Res Part E: Logist. Transp. Rev.*, 2018, 376–391.

Glosten, L.R., Jagannathan, R. and Runkle, D.E., On the relation between the expected value and the volatility of the nominal excess return on stocks. *J. Finance*, 1993, **48**, 1779–1801.

Gulin, A., Dorogush, A.V. and Ershov, V, *CatBoost: Gradient Boosting with Categorical Features Support*, 2018, arXiv:1810.11363.

Gunnarsson, E.S., Isern, H.R., Kaloudis, A., Risstad, M., Vigdel, B. and Westgaard, S., Prediction of realized volatility and implied volatility indices using AI and machine learning: A review. *Int. Rev. Financ. Anal.*, 2024, **93**, 103221.

Hansen, P.R., A test for superior predictive ability. *J. Bus. Econ. Stat.*, 2005, **23**, 365–380.

Hansen, P.R. and Lunde, A., A forecast comparison of volatility models: does anything beat a GARCH(1,1)? *J. Appl. Econom.*, 2005, **20**, 873–889.

Hansen, P., Lunde, A. and Nason, J., Choosing the best volatility models: The model confidence set approach. *Oxf. Bull. Econ. Stat.*, 2003, **65**, 839–861.

Hansen, P.R., Lunde, A. and Nason, J.M., The model confidence set. *Econometrica*, 2011, **79**, 453–497.

Herrera, A.M., Hu, L. and Pastor, D., Forecasting crude oil price volatility. *Int. J. Forecast.*, 2018, **34**, 622–635.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and Kolehmainen, M., Methods for imputation of missing values in air quality data sets. *Atmos. Environ.*, 2004, **38**, 2895–2907.

Kavussanos, M.G., Price risk modelling of different size vessels in the tanker industry using Autoregressive Conditional Heteroskedastic (ARCH) models. *Logist. Transp. Rev.*, 1996, **32**, 161.

Kavussanos, M.G., The dynamics of time-varying volatilities in different size second-hand ship prices of the dry-cargo sector. *Appl. Econ.*, 1997, **29**, 433–443.

Kavussanos, M.G. and Nomikos, N.K., Constant vs. time-varying hedge ratios and hedging efficiency in the BIFFEX market. *Transp. Res. Part E Logist. Transp. Rev.*, 2000, **36**, 229–248.

Kavussanos, M.G. and Visvikis, I.D., Market interactions in returns and volatilities between spot and forward shipping freight markets. *J. Bank. Financ.*, 2004, **28**, 2015–2049.

Kohavi, R. and John, G., Wrappers for feature subset selection. *Artif. Intell.*, 1997, **97**, 273–324.

Laurent, S., Rombouts, J.V.K. and Violante, F., On the forecasting accuracy of multivariate GARCH models. *J. Appl. Econom.*, 2011, **26**, 934–955.

Laurent, S., Rombouts, J.V. and Violante, F., On the forecasting accuracy of multivariate GARCH models. *J. Appl. Econom.*, 2012, **27**, 934–955.

Lipton, Z., Kale, D. and Wetzel, R, Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series. *CoRR*, 2016, **abs/1606.04130**.

Little, R.J. and Rubin, D.B., *Statistical Analysis with Missing Data*, 793, 2019 (John Wiley & Sons).

Liu, J., Li, Z., Sun, H., Yu, L. and Gao, W., Volatility forecasting for the shipping market indexes: an AR-SVR-GARCH approach. *Maritime Policy Manag.*, 2022, **49**, 864–881.

Moritz, S. and Bartz-Beielstein, T., imputeTS: time series missing value imputation in R. *R J.*, 2017, **9**, 207–2018.

Nelson, D.B., Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 1991, **59**, 347–370.

Pan, Z., Wang, Y., Wu, C. and Yin, L., Oil price volatility and macroeconomic fundamentals: A regime switching GARCH-MIDAS model. *J. Empir. Finance*, 2017, **43**, 130–142.

Patton, A., Volatility forecast comparison using imperfect volatility proxies. *J. Econom.*, 2011, **160**, 246–256.

Reddy, S., Akashdeep, S., Harshvardhan, R. and Kamath, S., Stacking deep learning and Machine learning models for short-term energy consumption forecasting. *Adv. Eng. Inform.*, 2022, **52**, 101542.

Sax, C. and Steiner, P., Temporal disaggregation of time series. *R J.*, 2013, **5**, 80–87.
Schnaubelt, M, A comparison of machine learning model validation schemes for non-stationary time series data. FAU Discussion Papers in Economics 11/2019, Nürnberg, 2019a.
Schnaubelt, M, A comparison of machine learning model validation schemes for non-stationary time series data. Technical report, FAU Discussion Papers in Economics, 2019b.
Stopford, M., *Marine Economics*, 2009 (Routledge).
Timmermann, A., *Handbook of Economic Forecasting*, 2013 (Newnes: Elsevier/North Holland).
Wang, Y. and Wu, C., Forecasting energy market volatility using GARCH models: Can multivariate models beat univariate models? *Energy Econ.*, 2012, **34**, 2167–2181.
Wang, L., Ma, F., Liu, J. and Yang, L., Forecasting stock price volatility: New evidence from the GARCH-MIDAS model. *Int. J. Forecast.*, 2020, **36**, 684–694.
Wang, X., Hyndman, R.J., Li, F. and Kang, Y., Forecast combinations: An over 50-year review. *Int. J. Forecast.*, 2023, **39**, 1518–1547.
Wujek, B., Hall, P. and Günes, F., *Best Practices for Machine Learning Applications*, 2016 (Las Vegas, NV: SAS Institute Inc).
Xu, J., Yip, T. and Marlow, P., The dynamics between freight volatility and fleet size growth in dry bulk shipping markets. *Transp. Res. Part E: Logist. Transp. Rev.*, 2011, **47**, 983–991.
Xu, L., Zou, Z. and Zhou, S., The influence of COVID-19 epidemic on BDI volatility: An evidence from GARCH-MIDAS model. *Ocean Coast. Manag.*, 2022, **229**, 106330.

## Appendices

### Appendix 1. Composition of baltic freight rates

Table A1. The composition of the different Baltic Exchange Average Trip-Charter rates are expressed by the routes and corresponding weights.

| Series | Routes Number | Description | Routes (weights) |
|---|---|---|---|
| Handysize | HS1_38 | Skaw-Passero trip to Rio de Janeiro-Recalada | 12.5% |
| Average 7 Trip Charter | HS2_38 | Skaw-Passero trip to Boston-Galveston | 12.5% |
| (HA7TC) | HS3_38 | Rio de Janeiro-Recalada trip to Skaw-Passero | 12.5% |
| | HS4_38 | US Gulf trip via US Gulf or north coast South America to Skaw-Passero | 12.5% |
| | HS5_38 | South East Asia trip to Singapore-Japan | 20% |
| | HS6_38 | North China-South Korea-Japan trip to North China-South Korea-Japan | 20% |
| | HS7_38 | North China-South Korea-Japan trip to southeast Asia | 10% |
| Supramax | S1B_58 | Canakkale trip via Med or Bl Sea to China-South Korea | 5% |
| Average 10 Trip Charter | S1C_58 | US Gulf trip to China-south Japan | 5% |
| (SA10TC) | S2_58 | North China one Australian or Pacific round voyage | 20% |
| | S3_58 | North China trip to West Africa | 15% |
| | S4A_58 | US Gulf trip to Skaw-Passero | 7.5% |
| | S4B_58 | Skaw-Passero trip to US Gulf | 10% |
| | S5_58 | West Africa trip via east coast South America to north China | 5% |
| | S8_58 | South China trip via Indonesia to east coast India | 15% |
| | S9_58 | West Africa trip via east coast South America to Skaw-Passero | 7.5% |
| | S10_58 | South China trip via Indonesia to south China | 10% |
| Panamax | P1A_82 | Skaw-Gib transatlantic round voyage | 25%, |
| Average 5 Trip Charter | P2A_82 | Skaw-Gib trip HK-S Korea incl Taiwan | 10% |
| (PA5TC) | P3A_82 | HK-S Korea incl Taiwan 1 Pacific round voyage | 25% |
| | P4_82 | HK-S Korea incl Taiwan trip to Skaw-Gib | 10% |
| | P6_82 | Dely Spore round voyage via Atlantic | 30% |
| Capesize | C8_14 | Gibraltar/Hamburg transatlantic round voyage | 25% |
| Average 5 Trip Charter | C9_14 | Continent/Mediterranean trip China-Japan | 12.5% |
| (CA5TC) | C10_14 | China-Japan transpacific round voyage | 25% |
| | C14 | China-Brazil round voyage | 25% |
| | C16 | Revised backhaul | 12.5% |

All weighted averages are scaled by a segment-specific factor to obtain the average TC freight rates (Baltic Exchange 2023).
Note: Detail of routes and vessel details can be found on Baltic Exchange website: https://www.balticexchange.com/en/data-services/market-information0/indices.html. The numerical value following an abbreviation (if labeled) denotes the typical tons deadweight (dwt) capacity (in '000s) of the ship type.

## Appendix 2. Hyperparameter tuning range

Table A2. Range of hyperparameters considered for CatBoost.

| Parameter | Range |
|---|---|
| Iterations[a] | 50, 100, 150, 200 |
| Max depth[b] | 4, 6, 10 |
| Learning rate[c] | 0.001, 0.01, 0.05, 0.1 |
| L2 regularization strength[d] | 1, 5, 9 |
| Random strength[e] | 1, 5 |

[a]Number of trees.
[b]Maximum depth of each tree.
[c]Step size shrinkage.
[d]Strength of L2 norm regularization (log-scale).
[e]Randomness when choosing splits.

Table A3. Range of hyperparameters considered for Random Forest.

| Parameter | Range |
|---|---|
| Estimators[a] | 10, 20, 50, 100, 200, 500 |
| Max depth[b] | 'None', 3, 5, 10 |
| Min samples split[c] | 2, 3, 5, 7, 10 |
| Min samples leaf[d] | 1, 2, 3, 4 |
| Max features[e] | 'sqrt', 'log2', 1, 3, 5, 7 |

[a]Number of trees.
[b]'None' means max depth not constrained.
[c]Minimum number of samples required to split an internal node.
[d]Minimum number of samples required to be at a leaf node.
[e]The number of features to consider when looking for the best split.

Table A4. Range of hyperparameters considered for XGBoost.

| Parameter | Range |
|---|---|
| Estimators[a] | 10, 20, 50, 100, 200, 500 |
| Learning rate[b] | 0.001, 0.01, 0.1, 0.2 |
| Max depth[c] | 3, 5, 7, 10 |
| Subsample[d] | 0.8, 0.9, 1.0 |

[a]The number of boosting rounds/trees.
[b]Step size shrinkage.
[c]Maximum depth of each tree.
[d]Fraction of rows sampled for each boosting round.