Vol. 3, Iss. 1 (2026) 48-54, DOI: 10.61552/JMES.2026.01.005

JMES

Journal of Management and Engineering Sciences

www.jmes.aspur.rs

# Distribution-Preserving Data Augmentation for Ship Fuel Consumption Prediction under Limited Logbook Data: A Bunker Vessel Case Study

Ye Si Thu Aung[a,*], Le Van Diem[a], Tran Hong Ha[a]

[a] *Vietnam Maritime University, Hai Phong, Vietnam.*

Keywords:

*Bunker ship*
*Data Augmentation*
*Fuel Consumption Prediction*
*Logbook Data*
*Machine Learning*

* Corresponding author:

*Ye Si Thu Aung*
*E-mail:*
*yesithuaung3000@gmail.com*

*Received: 15.12.2025.*
*Revised: 07.02.2026.*
*Accepted: 10.02.2026.*

A B S T R A C T

*Fuel consumption prediction in shipping is often a problem because of limited data. This study looks at how to predict fuel consumption using a method. With two years of records from a bunker vessel for training, only with 76 records before splitting. The things that affected fuel consumption were: how far the vessel traveled, how fast it went how cargo it carried, the trim of the vessel the wind speed and the height of the waves. Fuel consumption prediction is very important in shipping. We need to find a way to make good predictions, with the data we have. The baseline ensemble models, such as Gradient Boosting and Random Forest and XGBoost were trained on the dataset. These models gave the performance $R^2$ lower than 0.64. To this issue, we augmented the training dataset from 61 to 1000 with distributional distortion methods that are reliable with similar characteristics of the original dataset. The trained results achieved up to MAE of 0.027, RMSE of 0.050, and $R^2$ of 0.995. This indicates that, under limited-data, carefully constrained augmentation can recover predictive structure without introducing distributional distortion, enabling reliable fuel prediction.*



## 1. INTRODUCTION

Fuel consumption remains the dominant operational cost and emissions driver in commercial shipping. Regulatory instruments such as EEXI and CII convert this operational variable into a compliance constraint, not merely an efficiency target [1, 2]. In practice, however, fuel consumption is rarely observed under controlled conditions. It is produced by coupled effects of speed, loading, trim, and environment, recorded inconsistently across voyages, and documented primarily through logbooks rather than continuous sensors on many vessels [3]. The predictive problem is therefore not one of model expressiveness alone, but of inference under sparse, irregular, and partially structured data.

Conventional physics-based approaches estimate fuel consumption by resolving resistance and

propulsion components under assumed operating states. These methods provide interpretable baselines but degrade when confronted with operational variability that is not explicitly parameterized, including fluctuating weather exposure, trim adjustments, and human decision-making [4]. Their applicability is further limited when required inputs are unavailable or coarsely recorded, as is common in logbook-derived datasets. Under such conditions, prediction error reflects missing structure rather than hydrodynamic uncertainty.

Data-driven models address this limitation by learning empirical relationships directly from operational records. Machine learning methods, particularly tree-based ensembles, have shown the capacity to capture nonlinear dependencies among speed, displacement proxies, and environmental factors without explicit physical formulation [5, 6]. Their effectiveness, however, is strongly conditioned on data volume and diversity. Most reported successes rely on large, sensor-rich datasets that are unavailable for a significant portion of the active fleet. When sample size collapses to tens of observations, model variance dominates, and apparent underperformance becomes indistinguishable from data insufficiency [7].

This constraint defines the core problem addressed in this study. The objective is not to introduce a new predictive algorithm, but to examine how predictive structure can be recovered when only small-scale logbook data are available. Data augmentation offers one possible intervention, but naïve resampling or noise injection risks distorting operational distributions and violating domain constraints. For fuel consumption, even small distributional shifts can yield physically implausible predictions when extrapolated [8]. Augmentation must therefore be constrained, distribution-aware, and explicitly bounded by operational limits. The effectiveness of such augmentation should be evaluated not by in-sample fit alone, but by behavior on genuinely unseen voyages.

This study has done a simple but distribution-preserving augmentation technique by expanding the training set from 61 to 1000 data entries while maintaining covariance structure and physical bounds. The analysis focuses on whether data augmentation recovers stable predictive structure without introducing artefactual accuracy.

The results clarify how, under limited data conditions, constrained augmentation can shift model behavior from low performance to operationally high-reliable performance in ship fuel consumption prediction.

## 2. RESEARCH METHODOLOGY

### 2.1 Problem Formulation

Let the operational dataset be defined as a finite set $D = \{(x_i, y_i)\}_{i=1}^{N}$ ; where each record corresponds to a voyage segment extracted from ship logbooks. The feature vector $x_i \in \mathbb{R}^6$ ; contains operational variables: distance run, vessel speed, cargo load, wave height, wind speed, and trim. The target variable $y_i \in \mathbb{R}$ ; denotes the corresponding fuel consumption rate expressed in tons per day.

**Table 1.** Sample operational logbook records.

| Distance Run (NM) | Vessel Speed | Cargo Load | Wave Height (m) | Wind Speed (knots) | Trim | Fuel Consumption |
|---|---|---|---|---|---|---|
| 226 | 9.40 | 9417 | 3.25 | 18.4 | -0.02 | 10.50 |
| 257 | 10.70 | 9417 | 3.25 | 24.3 | -0.02 | 10.48 |
| 219 | 9.10 | 9417 | 3.25 | 24.3 | -0.02 | 10.50 |
| 236 | 9.70 | 2340 | 1.88 | 18.4 | -2.50 | 9.79 |
| 248 | 10.00 | 2340 | 1.88 | 18.4 | -2.50 | 9.80 |
| 224 | 9.33 | 7000 | 0.88 | 12.7 | -0.94 | 9.96 |
| 246 | 8.90 | 7000 | 0.88 | 7.8 | -0.94 | 10.21 |
| 232 | 10.00 | 6600 | 0.30 | 7.8 | -0.40 | 9.58 |
| 186 | 10.30 | 2500 | 0.05 | 4.9 | -2.10 | 7.35 |
| 258 | 10.70 | 2100 | 0.88 | 12.7 | -2.40 | 9.80 |

The prediction task is formulated as supervised regression: $\hat{y} = f(x)$ ; where $f(\cdot)$ is a learned mapping that estimates fuel consumption under a given operational state. No temporal ordering is assumed. Records are treated as conditionally independent samples due to irregular logging intervals and incomplete voyage continuity.

The central constraint is dataset size of 76 rows and 7 columns as shown in Table 1, the operating space is sparsely sampled, and conventional asymptotic assumptions underlying statistical learning do not apply. Model performance is therefore governed primarily by variance and sensitivity to sampling noise rather than representational capacity.

### 2.2 Data Validation and Cleaning

Prior to modeling, the raw dataset was subjected to automated validation using a Python-based inspection pipeline. The process targeted structural and semantic consistency rather than exploratory inference. Column names were standardized, non-numeric tokens removed, and mixed-format fields resolved. Range constraints were enforced for physically bounded variables, including wind direction, wind force, and fuel consumption, with violations logged for inspection.

Duplicate records were removed, missing values imputed using median statistics for numeric variables, and outliers adjusted using an interquartile-range-based winsorization procedure. This approach suppresses extreme leverage points while retaining rank structure within each variable. After cleaning, all variables were numeric, non-null, and bounded within operationally plausible ranges. The resulting dataset preserves empirical variability while eliminating artifacts introduced by logging inconsistencies.

### 2.3 Dataset Partitioning

To evaluate generalization under limited data, the cleaned dataset was partitioned into training and testing subsets using random sampling with a fixed seed to ensure reproducibility. Eighty percent of the records were allocated to training, with the remaining twenty percent reserved for testing. The split was performed once and held fixed across all baseline experiments to prevent evaluation drift.

Given the small sample size, no cross-validation was applied. Repeated resampling would have reduced the effective training set and amplified variance effects. Instead, emphasis was placed on consistency between training behavior and performance on genuinely unseen records.

### 2.4 Baseline Models

Three ensemble-based regression models were selected as baselines: Gradient Boosting Regressor, Random Forest Regressor, and Extreme Gradient Boosting. These models were chosen not for novelty, but for their known robustness under nonlinear, low-dimensional settings. All models were trained using identical feature sets and a uniform preprocessing pipeline consisting of median imputation. No hyperparameter tuning was performed at this stage in order to isolate the effect of data regime rather than optimization effort as shown in Fig. 1.

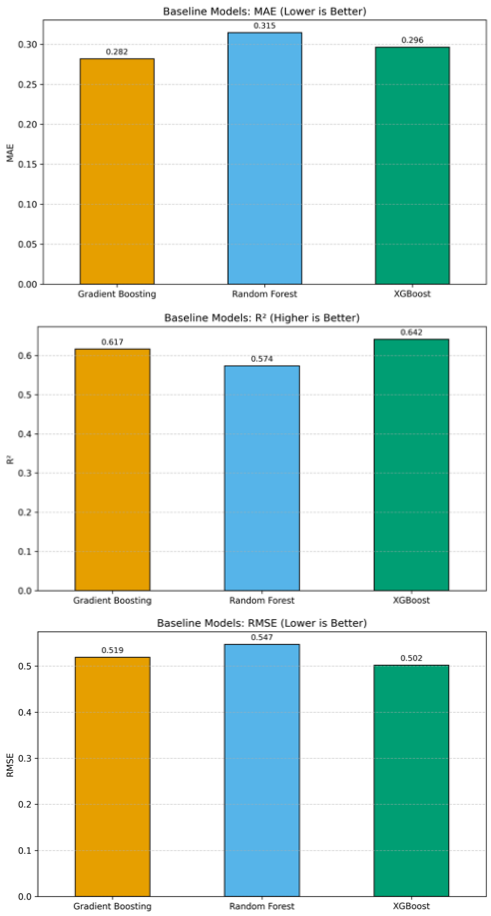


**Fig. 1.** The MAE, $R^2$, EMSE of three models trained with original dataset.

In Table 2, performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination ($R^2$). These metrics jointly characterize absolute deviation, sensitivity to large errors, and explained variance, which is necessary under small-sample conditions where single metrics can be misleading.

**Table 2.** The MAE, $R^2$, EMSE of three models trained with original dataset.

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| **XGBoost** | 0.2963 | 0.5017 | 0.6416 |
| **Gradient Boosting** | 0.2818 | 0.5189 | 0.6166 |
| **Random Forest** | 0.3147 | 0.5473 | 0.5736 |

## 2.5 Distribution-Preserving Data Augmentation

To address instability arising from limited sample size, a constrained data augmentation procedure was introduced. The training subset was expanded via bootstrap resampling with replacement, followed by the injection of Gaussian perturbations scaled to a fixed fraction of each feature's empirical standard deviation. Perturbation magnitude was set conservatively to avoid altering marginal distributions or introducing implausible operating states.

Post-perturbation, all features were clipped at physically meaningful bounds to enforce non-negativity and operational feasibility. The augmented samples were concatenated with the original training data to produce an expanded dataset of 1,000 records. This procedure preserves first-order statistics and approximate covariance structure while increasing coverage of the observed operating space.

## 2.6 Training on Augmented Data

The same baseline models were retrained on the augmented dataset using the identical preprocessing and evaluation protocol. Train–test splitting was repeated with the same random seed to ensure comparability with non-augmented results. Model artifacts, predictions, and performance metrics were stored for subsequent comparison.

The methodological focus is not on achieving maximal in-sample accuracy, but on evaluating whether augmentation stabilizes model behavior and improves predictive performance on held-out data without introducing distributional distortion.

## 2.7 Holdout Validation and Comparative Evaluation

Final evaluation was conducted using independent holdout records not involved in augmentation or model fitting. Predictions from models trained on original and augmented datasets were compared against observed fuel consumption values. Error distributions, percentage deviations, and global metrics were computed to assess calibration and generalization.

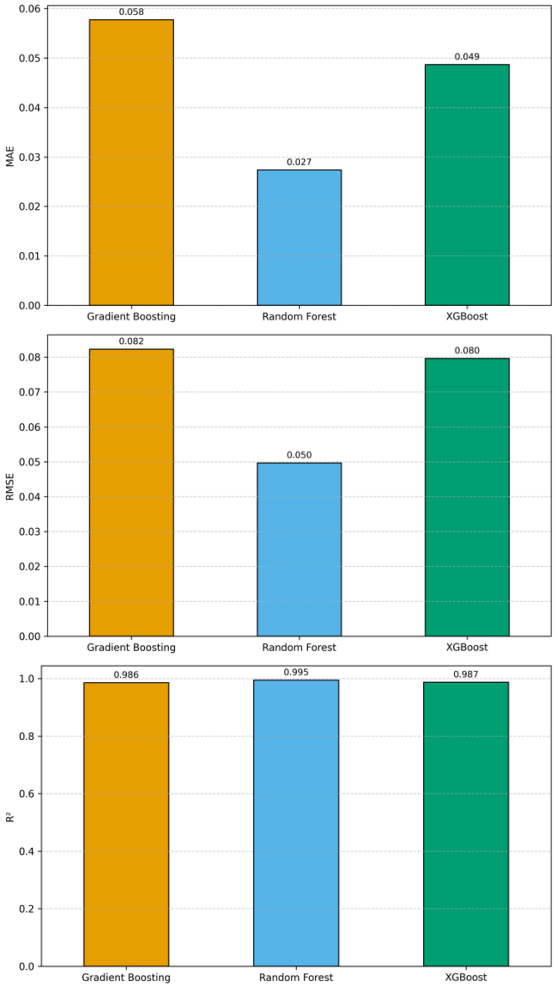


**Fig. 2.** The MAE, $R^2$, EMSE of new models trained with 1000 augmented data.

Comparative analysis emphasizes relative behavior between models trained under different data regimes. Improvements are interpreted as evidence of recovered predictive structure rather

than absolute performance gains in Fig. 2. This framing is necessary to distinguish genuine generalization from artefacts induced by synthetic data inflation.

## 3. RESULTS AND DISCUSSION

### 3.1 Baseline Performance on the Original Dataset

Baseline models trained on the original 76-record dataset exhibited constrained but interpretable performance. Among the three ensemble regressors, XGBoost achieved the highest explanatory power, with $R^2$=0.642, followed by Gradient Boosting ($R^2$=0.617) and Random Forest ($R^2$=0.574). Absolute error levels remained moderate across models, with MAE values clustered between 0.28 and 0.32 tons/day.

These results should not be interpreted as algorithmic inadequacy. With six input variables and sparse coverage of the operating space, the effective degrees of freedom are dominated by sampling variance. The observed ceiling in R2 reflects incomplete representation of operational regimes rather than systematic bias. Model rankings were stable across metrics, indicating that performance differences arise from variance reduction capacity rather than feature sensitivity.

Tree-based ensemble methods outperformed linear baselines under this regime, consistent with their ability to capture threshold effects and interaction structure without assuming functional form. However, none of the models demonstrated strong generalization margins at this scale, underscoring the limits imposed by data volume.

### 3.2 Effect of Distribution-Preserving Augmentation

Augmentation expanded the training set from 61 to 1,000 records while maintaining empirical distributions and physical bounds. Retraining on the augmented dataset produced a pronounced shift in model behavior. All three models exhibited substantial reductions in MAE and RMSE, accompanied by near-complete recovery of explained variance.

The Random Forest model achieved the strongest test performance, with MAE=0.027 tons/day, RMSE=0.050 tons/day, and $R^2$=0.995. XGBoost and Gradient Boosting followed closely, both exceeding $R^2$=0.985 as shown in Table 3. Error magnitudes decreased by an order of magnitude relative to the non-augmented baseline, indicating that variance suppression rather than marginal fitting drove the improvement.

**Table 3.** Performance comparison of machine learning models.

| Model | MAE | RMSE | $R^2$ |
|---|---|---|---|
| XGBoost | 0.049 | 0.080 | 0.987 |
| Gradient Boosting | 0.058 | 0.082 | 0.986 |
| Random Forest | 0.027 | 0.050 | 0.995 |

Importantly, these gains did not arise from distributional distortion. Summary statistics and feature ranges of the augmented dataset remained aligned with the original data, and predictions remained within operationally plausible bounds. The improvement therefore reflects stabilization of the learned mapping rather than artefactual overfitting.

### 3.3 Validation on Independent Holdout Voyages

Generalization was evaluated using held-out voyage records not involved in training or augmentation. When applied to this dataset, the baseline XGBoost model trained on the original data achieved $R^2$=0.816, with MAE=0.240 tons/day. The augmented Random Forest model improved performance across all metrics, reaching $R^2$=0.856 and MAE=0.172 tons/day.

Percentage-based errors followed the same pattern. Mean absolute percentage error decreased from 2.57% to 1.80%, and symmetric MAPE showed a comparable reduction as in table 4. These differences are operationally non-trivial given the narrow absolute range of daily fuel consumption observed in the dataset.

**Table 4.** Comparison of error percentage between baseline and augmented detection.

| Model | MAE | RMSE | $R^2$ | MAPE (%) | sMAPE (%) |
|---|---|---|---|---|---|
| Base | 0.240 | 0.300 | 0.816 | 2.57 | 2.54 |
| 1000 | 0.172 | 0.265 | 0.856 | 1.80 | 1.79 |

Direct comparison between the two prediction sets showed a mean symmetric percent difference of approximately 1.37%, indicating close calibration between models as shown in Fig. 3. The augmented model's advantage arises from reduced dispersion rather than systematic shift, suggesting improved robustness rather than altered bias.

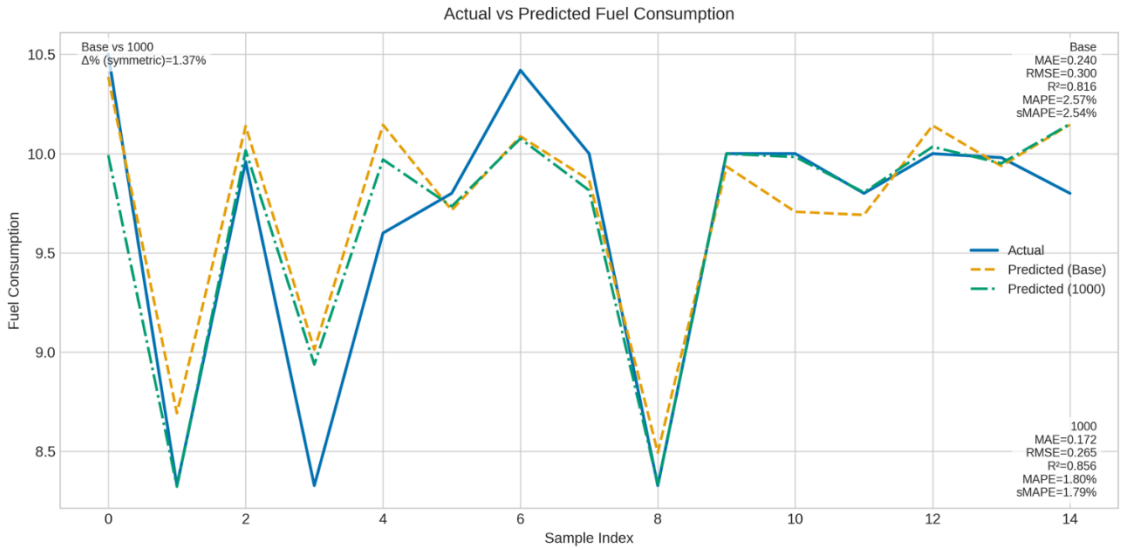


**Fig. 3.** Actual vs predicted fuel consumptions with unseen data.

### 3.4 Error Structure and Operational Consistency

Error distributions for both models were compact and centered near zero, with the augmented model exhibiting tighter concentration and fewer extreme deviations. Visual inspection of prediction-versus-actual plots confirmed close alignment with the identity line across the full operating range. Localized discrepancies persisted at the edges of the observed space, particularly under high wind and extreme trim conditions, but their magnitude was reduced after augmentation.

Predicted fuel consumption values responded monotonically to distance run and vessel speed, consistent with propulsion fundamentals. Secondary effects associated with trim and wind speed were preserved without amplification, indicating that augmentation did not introduce spurious sensitivities. This consistency is critical for operational use, where implausible response patterns undermine trust regardless of numerical accuracy.

### 3.5 Interpretation Under Small-Data Constraints

The results clarify the role of augmentation under logbook-scale data conditions. Without augmentation, ensemble models capture partial structure but remain variance-limited. With constrained augmentation, the same models recover stable mappings that generalize to unseen voyages without violating physical plausibility.

The improvement does not imply that synthetic data substitute for real observations. Rather, augmentation densifies the observed operating manifold, allowing ensemble methods to average over noise that would otherwise dominate fitting. The effectiveness of the approach depends on preserving covariance structure and enforcing domain constraints; unconstrained augmentation would likely inflate apparent accuracy while degrading operational validity.

From a methodological perspective, the findings indicate that data regime manipulation can be as consequential as model selection when observational capacity is limited. Under such conditions, predictive reliability is governed less by algorithmic sophistication than by how effectively the available data represent the operating space.

## 4. CONCLUSION

This study examined ship fuel consumption prediction under conditions where only small-

scale operational logbook data are available. Using two years of records from a bunker vessel, baseline ensemble models exhibited moderate predictive capability, with performance bounded by sample size rather than model structure. These results reflect a common operational reality: limited data coverage constrains inference more strongly than algorithm choice.

Introducing a distribution-preserving augmentation procedure altered this regime. By expanding the training set through constrained bootstrap resampling with covariance-scaled perturbations, model variance was substantially reduced without distorting empirical distributions or violating physical bounds. Ensemble models retrained on the augmented data achieved stable and accurate predictions, with improved generalization on independent holdout voyages. The observed gains are attributable to recovery of predictive structure rather than artificial inflation of fit.

The findings carry practical implications for fuel estimation in fleets lacking dense sensor instrumentation. Under such conditions, reliable prediction can be achieved without complex model architectures, provided that data handling preserves operational realism. Augmentation operates as a structural intervention on the learning problem, compensating for sparse coverage of the operating space while retaining domain consistency.

Several limitations remain. This study relies on a single vessel and a restricted set of operational variables. Although the augmentation procedure improves model stability, it does not introduce new physical regimes beyond those implicitly present in the original data. Extension to multi-vessel datasets would be required to assess transferability across different types of ships.

## REFERENCES

[1] "2023 IMO STRATEGY ON REDUCTION OF GHG EMISSIONS FROM SHIPS." ACCESSED: NOV. 14, 2025. [ONLINE]. AVAILABLE: HTTPS://WWW.IMO.ORG/EN/OURWORK/ENVIRONMENT/PAGES/2023-IMO-STRATEGY-ON-REDUCTION-OF-GHG-EMISSIONS-FROM-SHIPS.ASPX

[2] United Nations Conference on Trade and Development, Review of Maritime Transport 2024. 2024. Accessed: Oct. 16, 2025. [Online]. Available: https://unctad.org/system/files/official-document/rmt2024_en.pdf

[3] A. Fan, J. Yang, L. Yang, D. Wu, and N. Vladimir, "A review of ship fuel consumption models," Ocean Engineering, vol. 264, p. 112405, 2022, doi: 10.1016/j.oceaneng.2022.112405.

[4] R. Campbell, M. Terziev, T. Tezdogan, and A. Incecik, "Computational fluid dynamics predictions of draught and trim variations on ship resistance in confined waters," Applied Ocean Research, vol. 126, p. 103301, 2022, doi: 10.1016/j.apor.2022.103301.

[5] T. Uyanık, Ç. Karatuğ, and Y. Arslanoğlu, "Machine learning approach to ship fuel consumption: A case of container vessel," Transportation Research Part D: Transport and Environment, vol. 84, p. 102389, 2020, doi: 10.1016/j.trd.2020.102389.

[6] C. Papandreou and A. Ziakopoulos, "Predicting VLCC fuel consumption with machine learning using operationally available sensor data," Ocean Engineering, vol. 243, p. 110321, 2022, doi: 10.1016/j.oceaneng.2021.110321.

[7] S. Wang, B. Ji, J. Zhao, W. Liu, and T. Xu, "Predicting ship fuel consumption based on LASSO regression," Transportation Research Part D: Transport and Environment, vol. 65, pp. 817–824, 2018, doi: 10.1016/j.trd.2017.09.014.

[8] A. Jaswanth and D. D. Raju, "A MACHINE LEARNING MODEL FOR AVERAGE FUEL CONSUMPTION IN HEAVY VEHICLES," International Journal of Innovative Research in Technology, vol. 9, no. 12, pp. 1230–1239, 2023.