Parameter Uncertainty Quantification in Trans-Atlantic Vessel Voyage Time

by
Matthew Adiletta
and
Syed Mahdi

Submitted May 2026 in Partial Fulfillment of the Requirements for the Degree of Master of Applied Science in Supply Chain Management at the Massachusetts Institute of Technology

ABSTRACT

Maritime voyage duration is a critical but uncertain input to routing, scheduling, and downstream digital twin models for oil and gas transportation. This capstone develops a general framework for forecasting voyage duration under uncertainty using historical voyage data, vessel-status information, and weather proxies. The framework is demonstrated on Aframax tanker voyages along the Houston-to-Rotterdam trans-Atlantic crude oil route.

Records are cleaned through deduplication, multi-terminal merging, and timestamp validation. Each voyage is decomposed into five sequential legs using Kpler and Signal Ocean port-call and vessel-status data. These legs are origin waiting time, origin loading time, open-ocean transit time, destination waiting time, and destination discharge time. A nine-zone ERA5 significant wave height index is used as a weather proxy for the transit leg.

Five supervised machine learning models are developed, one per leg, evaluated across Random Forest, XGBoost, LightGBM, deep neural networks, and stacking ensembles with Bayesian hyperparameter tuning. We used mean absolute error (MAE) as a prediction error metric to evaluate these models. The best models achieve a MAE of 0.32 days for origin loading, 1.09 days for open-ocean transit, and 0.28 days for destination discharge. A Monte Carlo aggregator propagates empirical residual distributions across all five legs to produce a total voyage duration forecast of 25.5 days with a 90% confidence interval of [21.0, 32.5] days, enabling the sponsor to move from point-estimate voyage planning to probabilistic scheduling that can support downstream routing and optimization models.

Capstone Co-Advisor: Tim Russell
Title: Research Engineer, MIT Center for Transportation and Logistics
Capstone Co-Advisor: Thomas Koch
Title: Postdoctoral Associate, MIT Center for Transportation and Logistics

Table of Contents

1. Introduction

The project sponsor is a large, multinational oil and gas company with operations spanning the entire globe. They are a fully integrated supply chain company that handles exploration, drilling, transportation, refining, and final sales to customers. The company has a large global presence, with operations across many countries and a workforce of tens of thousands of employees (Sponsor Company, n.d.). The sponsor spends upwards of US $10 billion a year on transportation of chemicals, crude oil, and other natural resources.

1.1 Motivation

Maritime voyage duration is a critical input to routing, scheduling, and operational planning decisions. However, total voyage time can vary significantly due to factors such as weather conditions, port congestion, operational delays, and execution variability across different phases of a voyage. This variability makes voyage duration inherently uncertain, creating challenges for planning and optimization models that rely on accurate operational inputs.

Traditional forecasting approaches for voyage time planning typically provide point estimates without quantifying the uncertainty surrounding those predictions. However, downstream optimization models require probability distributions rather than single estimates to properly represent operational variability and risk. When this uncertainty is poorly characterized, the effectiveness of planning and optimization tools can be significantly degraded.

This project focuses on quantifying and communicating voyage-time uncertainty so that it can be more effectively incorporated into downstream decision models. Voyage duration, along with the variability in its underlying predictors, serves as a representative example of an uncertain forecast input to a broader voyage optimization program. Aframax tanker voyages are used as the case study for developing and demonstrating this uncertainty quantification framework.

1.2 Problem Statement and Key Questions

The key objective of the project is to perform statistical analysis of potential inputs in a forecast model, to better understand the impact that they have on the final time prediction. To do this we collected historical voyage data from the sponsor, fit probability distributions to key parameters (e.g., port delay, vessel speed, weather), and ran Monte Carlo (MC) simulations (discussed in Section 2.4) to evaluate how these uncertainties propagate through total voyage time. Sensitivity analysis then identified which parameters contributed most to variability. The main questions we answered are:

- What are the primary inputs that have an impact on a vessel's voyage time?
- What are the probability distributions associated with each of these inputs?
- How do the variations in these inputs affect the final voyage time forecasts?

1.3 Project Goals and Expected Outcomes

The project's overall goal is to improve forecasting accuracy of the sponsor's Trans-Atlantic vessel voyage time to support downstream optimization and planning models. Deliverables include a statistical model quantifying voyage-time uncertainty and a report on the analysis of its input parameters, enabling more reliable probabilistic inputs for scheduling and downstream models. In practice, the scope is to decompose total voyage time into five sequential legs, integrating Signal Ocean anchorage data with Kpler port-call records, and to apply per-leg machine learning models, producing a more granular uncertainty framework.

2. State of the Practice

Because this project aims to improve trans-Atlantic voyage-time forecasting and quantify planning uncertainty, this section reviews four themes central to any credible voyage-duration study: systematic reconstruction of AIS and commercial maritime datasets; goodness-of-fit methods for characterizing the stochastic behavior of voyage components; ERA5-derived significant wave height as a physically grounded transit-time driver; and the regression, machine learning, and simulation methods that translate cleaned data into calibrated uncertainty intervals.

2.1 Data Processing

Maritime voyage datasets derived from AIS are widely recognized as requiring extensive preprocessing. Zhang et al. (2022) emphasize that AIS data suffers from "noise, incorrect data, missing data, and packet loss during creation, encoding, transmission, and decoding." Liang et al. (2024) similarly document that "the original AIS data often suffers from unwanted noise (i.e., poorly tracked timestamped points for vessel trajectories) and missing ... data during signal acquisition, transmission, and analog-to-digital conversion," with this degradation posing "significant risks, including potential miscalculations in vessel collision avoidance systems" and other maritime operations. Prior research emphasizes that raw tanker-movement data rarely maps cleanly into single voyages (Wu et al., 2020) and therefore requires systematic reconstruction. Multiple studies advocate for statistical methods to identify and remove degraded data points (Liang et al., 2024; Wu et al., 2020; Zhang et al., 2022).

A recurring challenge is the presence of duplicate entries. AIS datasets often contain "duplicate entry of some data," because multiple receivers log the same transmission, requiring explicit deduplication steps (Zhang et al., 2025). Commercial services such as Kpler already pre-process and resolve most of these raw-AIS issues for commercial analysis, delivering cleaner voyage-level records to end users. However, further handling and formatting is required to tailor the dataset for our machine learning model — particularly around terminal aggregation and voyage reconstruction, as detailed below.

A second challenge is inconsistent port event recordings. AIS data "must be preprocessed to minimize event omissions or incorrect event recordings" (Zhang et al., 2022), requiring "uncertain reasoning" to handle ambiguous berthing patterns (Wu et al., 2020). For example, in Kpler, vessels calling Houston or Rotterdam appear under multiple terminal names, requiring heuristic grouping.

Finally, many tanker voyages involve multi-stage cargo handling, necessitating trajectory reconstruction (Liang et al., 2024). Kpler expresses these as multiple rows under a single Voyage ID, which must be merged to reconstruct full discharge sequences.

Across the literature, the consensus is that voyage reconstruction requires domain-specific heuristics, careful merging of multi-stage events, and robust timestamp consistency checks before any statistical analysis or modeling can be conducted.

2.2 Distribution Classification

A common method for assigning theoretical probability distributions to transportation data is the Kolmogorov-Smirnov (KS) test, a nonparametric approach particularly effective at detecting deviations near the central tendency of empirical distributions (Press & Teukolsky, 1988). Harsha and Mulangi (2021) compared the KS, Chi-squared, and Anderson-Darling tests in the context of transit travel time variability, finding that the Chi-squared test is sensitive to sample size and the Anderson-Darling test gives greater weight to tail deviations. For datasets with sparse tails and limited observations, the KS test offers the most robust fit assessment, making it well-suited to maritime voyage-time data where extreme values are relatively rare.

2.3 Quantifying Weather as a Variable

The maritime engineering literature shows that higher significant wave height (H_s) is associated with increased wave-induced resistance, reduced attainable vessel speed, higher fuel consumption, and potentially longer transit times (Faltinsen, 1990; IMO, 2017; Adland & Jia, 2018). For this reason, H_s is used in this study as a proxy for sea-state severity and weather-related transit disruption.

Sea-state parameters such as H_s, mean wave period, and swell characteristics originate from the wave-energy spectrum, which describes how ocean-surface variance is distributed across wave frequencies and directions (Komen et al., 1994; Holthuijsen, 2007). Because energies from unrelated wave systems simply add together, the integrated spectrum, the zeroth moment m_o, captures the combined effect of local wind waves and remote swell, both of which affect vessel motions. The commonly used relationship, $H_s = 4\sqrt{m_0}$ makes H_s a robust statistical measure of total sea-state severity, and it is the standard metric used in naval architecture, performance assessment, and weather routing.

For global ocean-state characterization, modern studies rely on ERA5, the fifth-generation European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric–wave reanalysis. ERA5 reconstructs waves every hour using the third-generation Wave Model (WAM) spectral wave model, assimilating buoy, altimetry, and wind observations to produce spatially consistent global fields at 0.25° resolution (Hersbach et al., 2020; WAMDI Group, 1988). Each 0.25° × 0.25° grid cell contains an independent time series of wave parameters, enabling detailed reconstruction of the environmental conditions faced along a given shipping corridor. ERA5 H_s fields therefore provide a physically grounded and operationally relevant proxy for voyage-time variability.

2.4 Data Analysis

Beyond distribution fitting, regression models have been widely applied to predict vessel arrival times in global supply chains. Pellemans et al. (2025) demonstrated the utility of regression and classification models in estimating expected time of arrival (ETA) using historical AIS data and vessel-specific features. Regression provided continuous predictions of travel times, enabling fine-grained operational planning for port scheduling and resource allocation. This work illustrates how regression can complement distributional analysis by offering explanatory insights into the drivers of variability.

Finally, uncertainty modeling has emerged as a critical component of travel time prediction. Naik et al. (2018) advanced this area by applying resampling methods, including bootstrapping, to estimate uncertainty in large probe datasets. They emphasized the role of Monte Carlo simulation in validating uncertainty estimates when true distributions are unknown, particularly in contexts where data scarcity limits the reliability of direct statistical inference. Monte Carlo simulation is a process by which “a model of possible results [is generated] by leveraging a probability distribution for any variable that has inherent uncertainty” (IBM). It helps to show the range of all the possibilities in a system and quantifies these ranges into confidence ranges. The work of Naik et al. (2018) highlights the importance of simulation-based approaches in extending the analytical power of limited datasets.

Taken together, these studies define the state of practice in travel time analysis: fitting probability distributions to empirical data, applying regression models for explainability and prediction, and employing resampling or simulation methods to quantify uncertainty.

2.5 Machine Learning for Voyage Time Prediction

Accurately forecasting voyage duration requires capturing nonlinear relationships between operational, temporal, and environmental inputs that standard regression approaches cannot represent. Pellemans et al. (2025) show that advanced machine learning models, including gradient-boosted trees and other

nonlinear approaches, achieve strong performance in vessel arrival time prediction when using operational, temporal, and environmental features; consistent with broader literature indicating that such models typically outperform linear regression (OLS) in this context. Their work motivates the adoption of Random Forest, XGBoost, and LightGBM as competing estimators in this framework.

Quantifying uncertainty around these predictions is equally important. Split-conformal prediction intervals, as formalized by Angelopoulos and Bates (2022), provide distribution-free uncertainty bounds with marginal coverage guarantees. This is a critical property for a sponsor relying on probabilistic intervals to inform downstream optimization models and accurately capture uncertainty, variability, and risk in voyage time planning.

3. Methodology

This study builds a probabilistic forecasting framework for Houston–Rotterdam Aframax crude oil voyages. Using commercial AIS data from Kpler and Signal Ocean, each voyage is decomposed into five sequential legs: origin waiting time, origin loading, open-ocean transit, destination waiting time, and destination discharge. A weather severity index derived from ERA5 wave height data augments the transit leg with environmental context. Five machine learning models (one per leg) are trained and selected across four model families, with the best performer per leg carried into an uncertainty quantification step. Leg-level uncertainty is quantified through split-conformal prediction intervals, and the underlying empirical residual distributions are propagated through Monte Carlo simulation to produce a full distribution of total voyage durations and percentile-based planning metrics.

3.1 Data Collection and Preparation

Voyage records were sourced from Kpler and reconstructed into 174 Houston-to-Rotterdam Aframax crude shipments spanning December 2015 to September 2025, through deduplication, multi-terminal merging, and timestamp validation. Figure 1 shows the resulting three-part decomposition (load time, sea-leg time, and discharge time) defined by four Kpler timestamps. This structure is refined in Section 3.4, where Signal Ocean vessel-status data split the load and sea-leg intervals into five modelable legs by isolating pre-port waiting time.

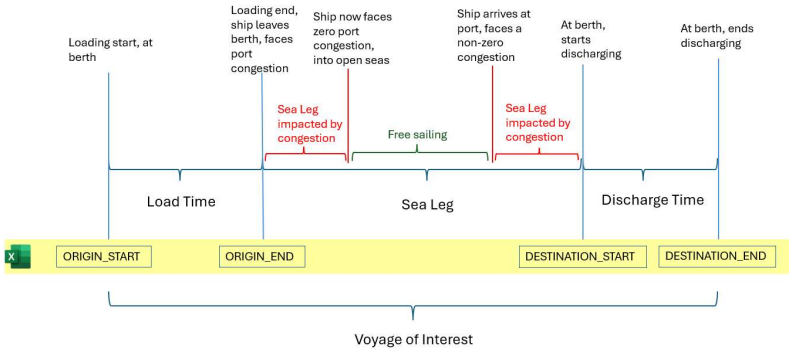

Figure 1: Voyage of Interest Primary Decomposition

3.2 Weather Proxy Construction

To quantify weather as a driver of transit time on the Houston-Rotterdam corridor, a zone-based sea-state severity index was constructed using ERA5 significant wave height (H_s). Nine geographic zones were defined along the route, covering the Gulf of Mexico departure, Florida Straits, Western Atlantic, North Atlantic storm track, and Rotterdam approaches (Figure 2). For each zone, monthly H_s was downloaded from ERA5 at 0.25° resolution and spatially averaged across all grid cells to produce a single time series per zone over 2010-2025 (Figure 3). Each zone's series was then normalized to a 0-1 index using P5-P95 scaling, where 0 represents unusually calm months and 1 represents historically severe months, preserving seasonality while preventing storm outliers from distorting the scale.

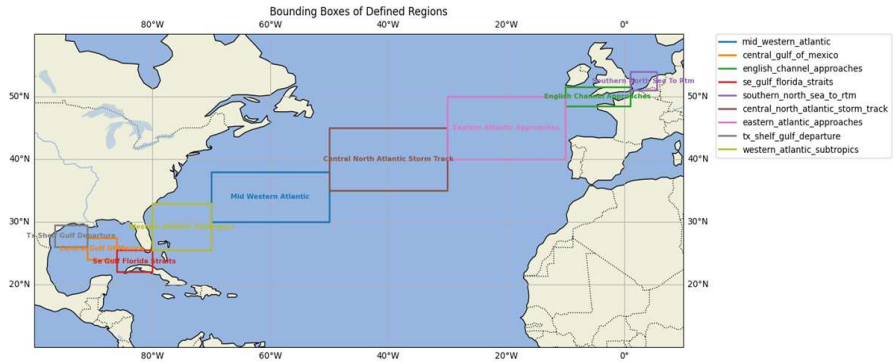

Figure 2: Defined Regions for Wave Height Assessment

Each voyage's overall weather severity index is a distance-weighted sum of the nine zone indices: $I_{voyage} = \Sigma\ (from\ z\ =\ 1\ to\ 9)[\ w_z * I_z]$ where w_z is the fraction of the voyage's sea-leg distance traversed in zone z, and I_z is that zone's normalized H_s index for the month of traversal. For voyages spanning a month boundary, zone indices were allocated proportionally by distance.

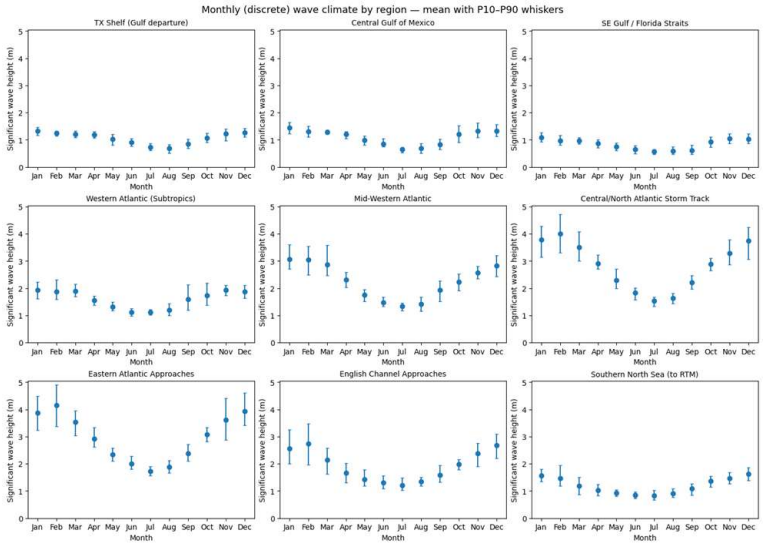

Figure 3: Sample Wave Climate by Region

3.3 Statistical Analysis

Initial analytical work combined distribution fitting, linear regression, and Monte Carlo simulation to characterize voyage-time variability across the four Kpler-derived legs. While this confirmed that seasonality, weather conditions, and port congestion were identifiable drivers of voyage duration, the linear framework explained only roughly half of the observed variation and could not capture the nonlinear interactions between vessel characteristics, weather, and port conditions that drive leg-level variability. These limitations motivated the machine learning architecture described in Section 3.5.

3.4 Signal Ocean Integration and Voyage Leg Engineering

The Kpler dataset provides four operational timestamps per voyage: ORIGIN_START, ORIGIN_END, DESTINATION_START, and DESTINATION_END. While these support load time, transit time, and discharge time calculations, they cannot distinguish between time spent waiting at anchorage versus time alongside the berth. To isolate waiting as a separable, predictable component, we integrated Signal Ocean daily vessel status updates, which record a vessel's state (loading, waiting to load, waiting to unload, or sailing) at each daily observation.

From these two sources, five voyage leg durations were engineered as follows. Origin waiting time (Leg 1) is derived from Signal Ocean's "waiting to load" status observations and represents the interval between a vessel's arrival in the Houston Ship Channel area and the Kpler-recorded ORIGIN_START event. Origin time (Leg 2) is the Kpler-defined loading duration: ORIGIN_END minus ORIGIN_START. Voyage time (Leg 3) is computed as the Kpler interval from ORIGIN_END to DESTINATION_START, minus the separately estimated destination waiting time, to isolate pure transit from confounding port-approach wait times. Destination waiting time (Leg 4) is derived from Signal Ocean's "waiting to unload" observations, representing the anchorage wait between vessel arrival at Rotterdam and the Kpler-recorded DESTINATION_START. Destination time (Leg 5) is the Kpler-defined discharge duration:

DESTINATION_END minus DESTINATION_START. Kpler serves as the primary source for voyage identity, cargo attributes, and port-call timestamps; Signal Ocean is used exclusively to engineer the waiting legs, providing vessel-state context that Kpler's port-call records cannot distinguish

Outlier filtering differs by leg type. Waiting legs apply a 99th percentile upper trim only, preserving zero-waiting observations and the scarce non-zero congestion cases while removing implausible extremes. Operational legs apply a bilateral 2.5th and 97.5th percentile trim because extreme values at either end would disproportionately influence regression training on a small dataset, pulling predictions away from the typical operational range. After filtering, waiting models use 172 voyages and operational models use 154–164 voyages. Table 1 summarizes the distribution of each leg.

Leg	Mean (days)	Std	% Zero	P90 (days)	Max (days)	n (samples)
Origin Waiting	2.6511	3.1287	27.01	6.1167	18.1667	172
Origin Time	1.9426	0.6817	0.00	2.8951	5.0000	164
Voyage Time	18.9527	2.8524	0.00	20.8802	41.7979	164
Destination Waiting	1.0266	2.2791	60.34	3.1333	15.1667	172
Destination Time	1.4913	0.4635	0.00	2.0775	3.0403	154

Table 1: Distribution of the Five Voyage Legs Durations After Filtering Outliers

3.5 Machine Learning Model Architecture

Five separate supervised models were developed, one per voyage leg. The choice of model architecture was determined by the statistical structure of each target. Leg 4 (destination waiting time) contains a majority of zero values, with 60% of voyages showing no measurable waiting time at Rotterdam. Leg 1 (origin waiting time) shows the opposite pattern, with 73% of voyages recording measurable congestion, consistent with the Houston Ship Channel's high utilization rate over the study period. Predictions are clipped at zero to enforce the physical constraint that waiting time cannot be negative. This approach avoids the classification instability that arises when the congested training subset is small and allows the regression to learn the full distribution including zeros directly.

For the three operational time legs (origin time, voyage time, and destination time), standard single-stage regression is appropriate because the targets are continuous with no structural zeros. All five models were developed under a consistent preprocessing pipeline: numerical features are median-imputed and StandardScaler-normalized; categorical features are one-hot encoded with frequency imputation for rare categories. Month is encoded as a sine term to capture seasonality. The voyage time model retains both sine and cosine month terms to preserve full cyclical representation, as the open-ocean transit is more sensitive to fine-grained seasonal weather patterns than the port legs.

Each leg's feature set was selected based on physical relevance, permutation importance screening, and collinearity analysis. Table 2 summarizes the final feature sets in plain operational terms; internal variable names are omitted. Features removed after screening include load port identity (100% Houston across all voyages, zero importance), discharge port identity (99% Rotterdam, zero importance), a

destination cargo weight field (0.89 correlation with cargo volume), vessel size from both waiting models (near-zero permutation importance), and port days waiting aggregates (superseded by the vessels-waiting count metric.)

Model	Features
Origin Waiting	terminal identity; cargo product grade; cargo volume; month of arrival (seasonal); number of vessels waiting at origin port; day of week of arrival
Origin Time	cargo product grade; month of loading; day of week of loading; cargo volume
Voyage Time	cargo volume; vessel year built; average laden sailing speed; month of departure (seasonal); Gulf of Mexico departure weather severity; open North Atlantic weather severity; Rotterdam approaches weather severity
Destination Waiting	discharge terminal identity; cargo product grade; cargo volume; month of arrival (seasonal); number of vessels waiting at destination port; day of week of arrival
Destination Time	cargo product grade; month of discharge; year of discharge; day of week of discharge; cargo volume; vessel year built

Table 2: Feature Sets for Each of the Five Voyage Leg Models

For each of the five voyage legs, four model types were trained and compared: Random Forest (RF), XGBoost (XGB), LightGBM, and a deep neural network (NN). Model configurations were chosen with the small dataset (131 training rows per operational leg) as the primary constraint. Every architectural choice prioritizes preventing overfitting over maximizing expressive capacity.

Baselines**:** All four model types were first run at standard configurations to establish a benchmark. Table 3 summarizes the key parameters for the construction of the models and the packages that supported them. Shared for simple recreation and future reference with an explanation of the rationale.

Model	Key Parameters	Design Rationale
Random Forest	100 trees, unconstrained depth	Sklearn defaults; serves as stable tree-based baseline
XGBoost	100 trees, depth 6, learning rate 0.3	XGBoost defaults; aggressive learning rate for baseline only
LightGBM	200 trees, 15 leaves, min 5 samples per leaf, learning rate 0.05	Conservative leaf-wise settings to prevent overfitting on small data
NN (port legs)	2 ReLU layers (64, 32), Adam, learning rate 1e-3, L2 alpha=1e-4, early stopping patience 20	Smaller architecture matched to 4–6 input features per port leg
NN (voyage time)	3 ReLU layers (128, 64, 32), same optimizer and regularization	Wider architecture to capture interactions among 8 features including weather composites

Table 3: Baseline and Tuned Model Configurations

The NN design choices (L2 regularization, early stopping, and adaptive learning rate) are particularly important at this sample size. Without them, a network would overfit the training set rapidly and fail to generalize about the held-out test rows.

Beyond the baseline configurations, two refinements were layered on. Hyperparameter tuning of the strongest single tree-based learner (XGBoost) and a stacking ensemble that blends the tree-based models. Each addresses a different limitation of the baselines. The first squeezes more performance from a single model, the second hedges against any one model's errors on a small training set.

Hyperparameter tuning: XGBoost was tuned independently per leg using Bayesian optimization (50 trials, 5-fold cross-validation) across eight parameters: tree count, depth, learning rate, subsample ratio, column subsampling, minimum child weight, and L1/L2 regularization. The search space was bounded conservatively to prevent overfitting. The shallower depth found for port legs (depth 3) reflects their simpler feature relationships; the deeper configuration for voyage time (depth 6) captures interactions among weather composites, speed, and cargo volume.

Stacking ensemble: RF, tuned XGBoost, and LightGBM were combined under a Ridge meta-learner trained on five-fold out-of-fold predictions. Ridge was chosen specifically because its L2 regularization prevents the ensemble from overcommitting to whichever base model happened to perform best on a small training fold and shrinking blends weights toward equality produces a more stable combination than an unconstrained meta-learner.

The best-performing model per leg was selected based on test-set MAE and carried forward to uncertainty quantification and Monte Carlo aggregation.

3.6 Uncertainty Quantification

Uncertainty is quantified at two levels. At the leg level, split-conformal prediction intervals (Angelopoulos & Bates, 2022) with α = 0.10 provide 90%-coverage symmetric intervals of the form $\hat{y} \pm \hat{q}$, where $\hat{q}$ is the quantile of absolute calibration residuals held out from training. For the waiting legs, the finite-sample corrected quantile level $(n+1)(1-\alpha)]/n$ is applied to ensure valid coverage guarantees on small calibration sets; for the operational time legs, the standard $(1-\alpha)$ quantile is used given the larger calibration pool. A coverage target of 90% (α = 0.10) was selected following standard practice in the conformal prediction literature (Angelopoulos & Bates, 2022), balancing interval informativeness against the risk of excluding operationally significant tail outcomes. At the voyage level, a Monte Carlo simulation draws $N = 10{,}000$ residual samples from each leg's out-of-fold empirical residual pool, adds them to that leg's mean prediction, clips at zero, and sums across all five legs. This approach respects the actual skewed residual distributions of the waiting legs rather than assuming normality and propagates uncertainty through the sequential leg structure without requiring strict independence between legs. The output is the complete distribution of simulated total voyage durations, from which percentile-based planning metrics are directly read.

4. Results and Discussion

This section presents results across six areas: model performance across all variants and legs, feature importance and interpretability, leg-level uncertainty distributions, validation of the Monte Carlo simulation against historical data, total voyage duration forecast, and weather proxy contribution.

Together these form a complete picture of what the framework predicts, how confidently, and which operational factors drive each leg.

4.1 Model Performance

Figure 4 shows MAE across all model variants and all five legs. Table 4 summarizes the best-performing model per leg. Mean absolute percentage error (MAPE) is used as an error metric to express the percentage difference between the actual and predicted value. For origin waiting is not reported due to divide-by-zero distortion. Destination waiting MAPE is computed on non-zero rows only (40% of voyages with measurable waiting). 90% CI width reflects the split-conformal interval on the best model.

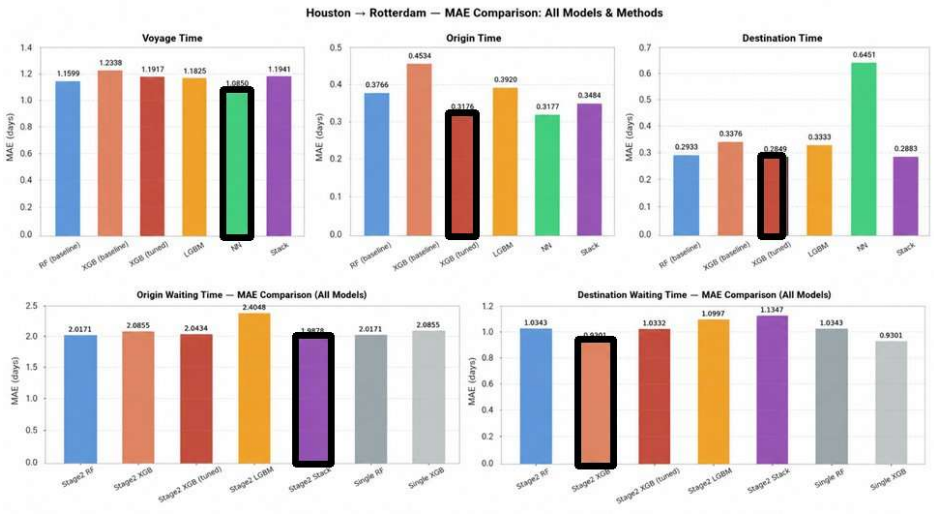

Figure 4: MAE Comparison Across All Model Variants and All Five Legs (RF, XGB, XGB Tuned, LGBM, NN, Stack)

Leg	Best Model	MAE (days)	MAPE	90% CI Width
Origin Waiting	Stacking Ensemble	1.988	—	±3.82 days
Origin Time	Neural Network	0.318	16.4%	±1.04 days
Voyage Time	Neural Network	1.085	5.6%	±2.11 days
Destination Waiting	XGBoost	0.930	74.2%*	±11.02 days
Destination Time	XGBoost (tuned)	0.281	20.3%	±0.68 days

Table 4: Final Model Performance by Voyage Leg

For voyage time, the neural network achieves the lowest MAE (1.085 days), outperforming the RF baseline (1.160 days) and tuned XGBoost (1.190 days). Its ability to capture weak nonlinear interactions among the weather composite features, which showed near-zero individual correlations ($|r| < 0.025$) with transit time but carry combined signal, likely explains this advantage. For origin time, the neural network and tuned XGBoost are tied (0.318 vs 0.318 days), with the neural network selected as the best model. Destination time is won by tuned XGBoost (0.281 days), with the neural network performing poorly on this leg (0.645 days), suggesting the discharge duration relationship is better captured by tree-based methods than a neural network on this small dataset. The stacking ensemble wins for origin waiting (1.988 days), with the meta-learner combining RF, XGBoost, and LightGBM signals that no single model captures alone. XGBoost wins for destination waiting (0.930 days).

Cross-validation fold variance provides important context for reliability. Origin time Cross-validation mean absolute error (CV MAE) ranged from 0.37 to 0.53 days across five folds; voyage time ranged from 0.76 to 1.42 days. Destination time was the most stable (0.22 to 0.30 days). This fold-level variance reflects sensitivity to which voyages fall in the test set, a consequence of the small dataset (n = 154–172 per leg).

4.2 Feature Importance and Model Interpretability

Permutation importance, which measures how much prediction error increases when a feature's values are randomly shuffled, was computed on the best model per leg. Figure 5 reports the results across all five legs.

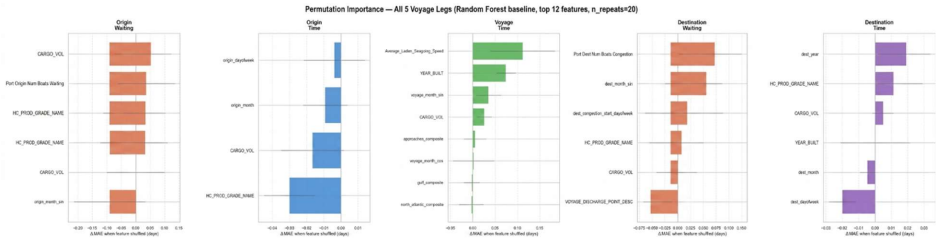

Figure 5: Permutation Importance for all Five Legs

For origin waiting, cargo volume emerges as the dominant predictor, followed by number of vessels waiting at the port, cargo product grade, and day of week of arrival. However, it is important to acknowledge that cargo volume and cargo grade may be acting as proxies for other operational factors (such as latent correlations with vessel type, terminal constraints, or time period) rather than direct causal drivers. Terminal identity has near-zero importance, indicating that berth assignment at Houston does not systematically predict waiting magnitude once other factors are controlled for.

For destination waiting, cargo volume is the strongest predictor, followed by month of arrival, cargo product grade, and number of vessels waiting at the port. Importance values are modest overall, confirming that Rotterdam waiting time is diffuse and difficult to attribute to specific operational factors.

For origin time, all features show near-zero or negative permutation importance. Cargo product grade shows a negative value, indicating that it adds noise rather than signal, making it a candidate for removal in future iterations. Month of loading and day of week contribute marginally. Like the origin leg, cargo volume here is likely acting as a proxy for latent operational constraints (such as shore tank

availability) rather than acting as a direct causal driver of delays. The weak importance profile across all origin time features is consistent with the low MAE (0.318 days) achieved by the neural network, which likely captures subtle patterns that permutation importance on the Random Forest baseline does not fully reflect.

For voyage time, average laden sailing speed dominates by a wide margin. As one would expect, this finding is essentially tautological since distance is roughly constant on a fixed Houston-Rotterdam route, making speed and time two sides of the same equation. The more interesting signal lies in the secondary predictors: vessel year built and the month terms (sine and cosine) follow, with the North Atlantic and Rotterdam approaches weather composites contributing modest positive signal. These features are what explain *why* speed varies from voyage to voyage, which is the more actionable question underneath. Notably, discharge terminal identity shows a small negative importance, indicating it adds noise to the transit model and should be removed from the voyage time feature set in future iterations.

For destination time, year of discharge is the top predictor, capturing longer-term trends in Rotterdam port throughput and capacity. Cargo product grade and month of discharge follow. Cargo volume shows negative importance, suggesting that it adds noise at the destination discharge stage and is also a candidate for removal. The year of discharge warrants further investigation to determine whether the observed effect reflects a multi-year trend in Rotterdam port capacity or is driven by a small number of anomalous years; connecting specific years to known operational changes at Rotterdam would strengthen the interpretability of this finding.

4.3 Leg-Level Uncertainty

Figure 6 shows the empirical residual distributions from out-of-fold predictions for each leg. The shape of each distribution, symmetric or skewed, narrow or long-tailed, matters because the Monte Carlo aggregator samples from these pools directly, so any asymmetry at the leg level propagates into the total voyage forecast.

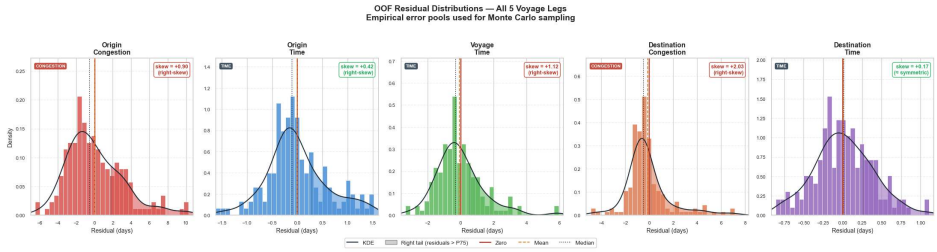

Figure 6: OOF Residual Distributions for all Five Legs

The waiting legs are heavily right-skewed (origin waiting skew +0.79, destination waiting skew +2.10), with most residuals clustered near zero but long positive tails extending to +10 days for origin waiting and +6 days for destination waiting. These tails represent voyages where the model substantially underestimated wait duration and are the primary driver of the upper tail in the total voyage distribution. Voyage time residuals also show moderate right skew (+1.25), reflecting the occasional influence of severe weather or routing delays on transit duration. The operational time legs (origin loading and destination discharge) show near-symmetric residual distributions (skew +0.34 and +0.22 respectively), centered close to zero on both sides, indicating well-calibrated regression with errors balanced across over- and under-prediction. The asymmetry in the waiting leg residual pools is why the

Monte Carlo total voyage P95 (32.5 days) sits 7.6 days above the P50 (24.9 days) while the P5 (21.0 days) is only 3.9 days below it.

4.4 Historical vs. Simulated Distributions

Figure 7 overlays the historical observed distribution of each leg against the Monte Carlo simulated distribution across the middle 95% of outcomes. Alignment is strong for the three operational legs (origin loading, open-ocean transit, and destination discharge), where MC medians are within 0.06 days of their historical counterparts and distribution shapes closely match. The waiting legs show broader divergence: origin waiting MC median (1.80 days) runs 0.47 days above the historical median (1.33 days), reflecting the right-skewed residual pool pulling simulated values toward higher leg duration outcomes. Destination waiting MC correctly reproduces the right skew and near-zero central tendency, though the sharp zero-mass spike in the historical data is smoothed in the simulation. Overall, the operational legs validate the residual sampling approach; the waiting leg divergence is consistent with the known difficulty of predicting rare, high-magnitude waiting events from a small calibration set.

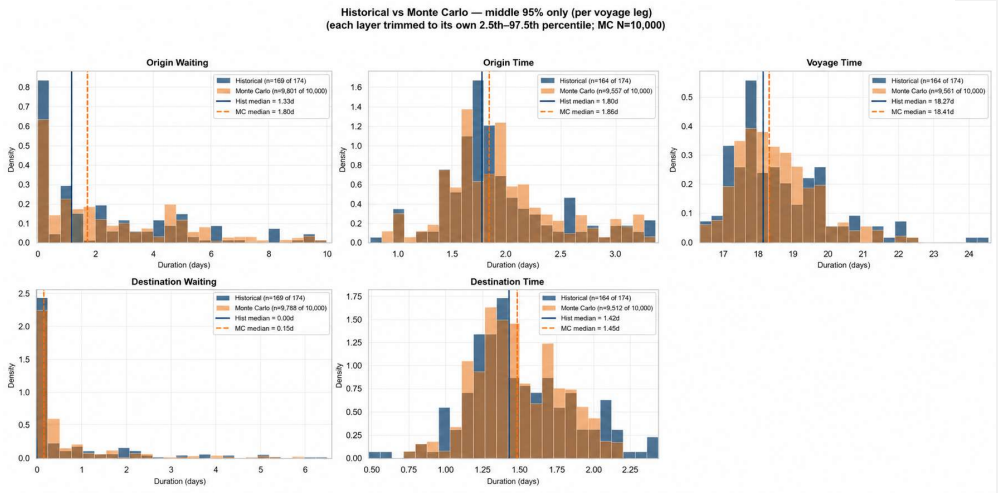

Figure 7: Historical Observed Distributions vs. Monte Carlo Simulated Distributions per Leg, Middle 95%

4.5 Total Voyage Duration

Figure 8 shows the sequential voyage legs with uncertainty whiskers spanning the 5th to 95th percentiles of each leg's MC distribution (the range covering 90% of simulated outcomes), alongside the total voyage MC distribution. The MC aggregation produces a mean total voyage duration of 25.53 days with a standard deviation of 3.58 days. The 90% confidence interval spans [21.03, 32.52] days. A coverage target of 90% ($\alpha = 0.10$) was selected following standard practice in the conformal prediction literature (Angelopoulos & Bates, 2022), balancing interval informativeness against the risk of excluding operationally significant tail outcomes. The P50 (median) of 24.90 days is slightly lower than the mean, reflecting the right-skewed contribution of the waiting legs. Table 5 presents the per-leg MC percentiles.

Voyage transit time accounts for approximately 74% of total mean duration and is the most predictable proportionally. Waiting accounts for approximately 12% of mean duration but contributes disproportionately to the upper tail of the MC P95 of 32.52 days versus P50 of 24.90 days represents a 7.6-day premium driven almost entirely by the right tail of origin waiting (P95 of 8.6 days versus a median of 1.9 days).

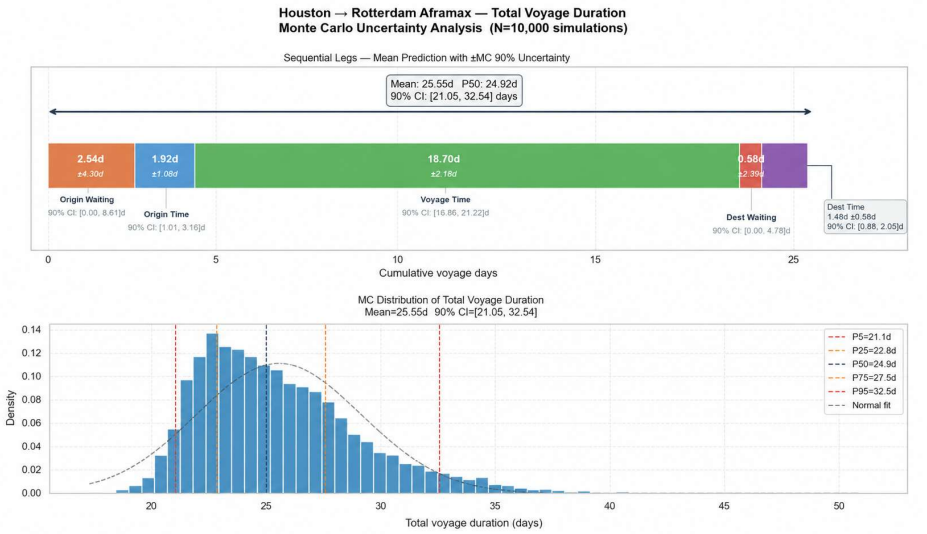

Figure 8: Per-leg duration with 90% Monte Carlo intervals (top) and the total voyage duration distribution (bottom), Houston → Rotterdam Aframax, N = 10,000 simulations.

Leg	Mean Pred (days)	MC P5 (days)	MC P50 (days)	MC P95 (days)
Origin Waiting	2.543	0.000	1.859	8.611
Origin Time	1.901	0.996	1.846	3.147
Voyage Time	18.698	16.864	18.410	21.219
Destination Waiting	0.583	0.000	0.168	4.783
Destination Time	1.469	0.875	1.446	2.039
TOTAL	25.194	21.03	24.90	32.52

Table 5. Per-leg and total voyage Monte Carlo percentiles (N = 10,000 simulations)

4.6 Weather Proxy Performance

The ERA5-based weather severity index contributes meaningfully to voyage time prediction despite near-zero individual correlations at the zone level ($|r| < 0.025$). Grouping nine individual zone indices into three geographic composites of Gulf departure, open North Atlantic, and Rotterdam approaches improved RF MAE from 1.242 to 1.160 days and XGB MAE from 1.341 to 1.234 days, a 6–8% gain. SHAP analysis of the voyage time model confirms that weather features rank among the top predictors: the mid-western Atlantic zone (0.218), western Atlantic subtropics (0.135), and Texas shelf Gulf departure (0.130) are secondary only to average laden sailing speed (0.478) and cargo volume (0.241). The near-zero linear correlations alongside meaningful SHAP importance confirm two things: sea-state severity is a genuine driver of transit time variability, and the nonlinear machine learning approach was necessary to extract that signal that linear regression would have missed it entirely.

4.7 Limitations

The results in this study should be interpreted as directional. The dataset of 154–172 voyages per leg, accumulated over a decade on a single route, is at the lower bound of what machine learning models require for reliable generalization. Cross-validation fold variance for voyage time ranged from 0.76 to 1.42 days across five folds. This nearly two-fold swing confirms that results are sensitive to which voyages happen to fall into the test set. A larger dataset would stabilize these estimates.

The waiting measurements carry a structural resolution problem. Signal Ocean records vessel status daily, meaning any waiting event shorter than 24 hours appears as zero. The zero rates in the data are therefore lower bounds on true waiting prevalence, not accurate counts. This directly affects the waiting models, which are trained on an incomplete picture of how often and how briefly vessels wait.

The split-conformal intervals for origin waiting achieved only 82.9% empirical coverage against the 90% target. The uncertainty bands for this leg are under-calibrated, meaning the stated interval is narrower than it should be. This is a consequence of the small, congested training subset and will improve as more positive examples accumulate.

Feature importance results should be interpreted as diagnostic rather than causal. Variables such as cargo volume and cargo product grade may be acting as proxies for operational conditions not directly observed in the dataset, including vessel utilization, terminal assignment, cargo parceling, or data-recording artifacts. These variables should be validated with operational experts before being used as decision rules.

For weather proxy, the regional categorization approach used may face generalization challenges beyond crude oil Aframax voyages. Clean product tankers, which exhibit higher load and discharge optionality and more varied routing, may require different zone definitions or alternative weather feature representations.

Finally, the framework covers one route, one vessel class, and one voyage type. It does not handle lightering, multi-port loading, or ship-to-ship transfers. These are not failures of the methodology — they are scope boundaries that define what the next phase of work needs to address.

5. Recommendations

This study develops and validates a probabilistic voyage time framework (i.e., five machine learning models estimating each voyage leg independently, combined through Monte Carlo simulation into a total voyage duration distribution) on the Houston-Rotterdam Aframax corridor. The results demonstrate that leg-level decomposition, machine learning regression, and conformal uncertainty

Commented [DA1]: Primary recommendation should be around future work vs "production ready". Or at least, the recommendation should be to go apply X mathematical techniques, rather than go use this specific prediction of 25days)

quantification together produce a more informative planning input than a single point estimate. The recommendations below address how the sponsor company could to act on the current results operationally and extend the methodology toward a fleet-wide capability.

5.1 Management Recommendations

This study demonstrates that voyage duration on the Houston-Rotterdam Aframax corridor can be decomposed into five predictable components and that uncertainty around each component can be quantified and aggregated into a total voyage distribution. The immediate recommendation is not to treat the specific outputs (a mean of 25.5 days, a 90% interval of [21.03, 32.52] days) as a final production forecast, but to instead use them as proof of concept that probabilistic voyage planning is achievable with the data sources and modeling approaches described here.

Three operational actions follow directly from the current results:

First, replace single-point ETA estimates with interval-based planning for the Houston-Rotterdam corridor. The P50 of 24.90 days is the base scheduling estimate; the P90 of 30.40 days is the appropriate upper bound for time-sensitive or contractually constrained decisions. Using a point estimate alone ignores the 7.6-day gap between median and P95 that waiting tail risk creates. Interval-based planning allows the sponsor to calibrate schedule buffers to an explicit risk tolerance, using P75 for routine scheduling and P95 for contractually sensitive decisions, rather than applying a fixed conservative estimate uniformly.

Second, treat origin waiting as a manageable variable rather than an uncontrollable one. Day of week of arrival is the most directly actionable predictor of Houston waiting time. This is where upstream scheduling flexibility exists, timing vessel arrivals to avoid historically high-waiting days of the week is a low-cost lever that does not require a deployed model to implement. Cargo volume and product grade also emerge as strong predictors, though these likely reflect latent operational patterns rather than direct causal relationships and warrant further validation before being used as planning inputs.

Third, focus on improving the underlying data integration before expanding the model. The sponsor should improve the Kpler-Signal matching layer and audit whether existing Signal Ocean speed and status fields are being fully extracted and retained for laden sea-leg days. Requesting hourly vessel status resolution from Signal Ocean rather than daily would further reduce ambiguity in waiting leg boundaries, as the current 24-hour floor obscures short-duration events. High-frequency AIS reconstruction should be considered selectively for validation or high-value lanes, not as the default global operating model.

Fourth, focus planning buffers on waiting legs rather than applying uniform buffers across the full voyage. Waiting accounts for only 12% of mean voyage duration but is responsible for most of the gap between P50 (24.9 days) and P95 (32.5 days). A uniform conservative buffer applied to the full voyage wastes buffer capacity on the predictable transit leg while underweighting the tail risk that drives scheduling failures.

5.2 Future Work

The primary contribution of this study is methodological: a replicable framework for decomposing voyage time into sequential legs, modeling each leg independently with machine learning, quantifying leg-level uncertainty through conformal prediction, and aggregating to a total voyage distribution through Monte Carlo simulation. The specific Houston-Rotterdam results are illustrative. The framework itself is what should be carried forward.

The natural next step is applying this framework to additional routes. The five-leg decomposition is route-agnostic; what changes per corridor is the training data, terminal features, and weather zone boundaries, not the architecture. The open-ocean transit leg offers the strongest standardization potential, as the ERA5 weather proxy and speed-based features are route-adaptable without requiring terminal-specific knowledge. For any origin-destination pair, the great-circle path can be divided into geographic zones, wave height extracted per zone, normalized, and weighted by voyage distance share.

Data volume is the binding constraint for fleet-wide scaling. This study required a decade of voyages on one of the sponsor's highest-frequency routes to reach 154 to 172 usable observations per leg. Shorter or less frequent routes will not have this history. The practical solution is pooling similar corridors, for example all US Gulf to Northwest Europe Aframax voyages, into shared training datasets with route-specific indicator features, trading some route-specific precision for network-wide generalizability.

Several methodological improvements are recommended for future iterations. First, leg boundaries should be anchored on a single consistent timestamping mechanism rather than aligned through temporal proximity heuristics between Kpler and Signal Ocean, which can introduce boundary misalignment for short-duration waiting events. Second, feature importance findings, particularly for cargo volume and product grade, should be validated with domain experts before being used as planning inputs, as these variables may be capturing latent correlations rather than direct causal effects. Third, the framework currently handles direct port-to-port voyages only; a production system must first classify voyage type and route lightering, multi-port load, and ship-to-ship transfers to appropriate leg architectures. Fourth, conformal calibration thresholds should be treated as living parameters and re-fitted on a rolling window of recent voyages, as port conditions and trade flows evolve over time.

For the sponsor, the value of this framework is the operating capability it demonstrates: probabilistic, leg-level voyage forecasting deployable across the fleet and consumable by downstream planning systems. That capability connects to three priorities.

Cost effectiveness: Uncertainty is concentrated in the waiting legs, which represent only 12% of mean voyage duration but drive the majority of worst-case schedule overruns. Replacing uniform buffers with leg-specific intervals reduces unnecessary slack on predictable transit days and concentrates buffer where tail risk lives, translating into fewer demurrage exposures, tighter commercial windows, and lower working capital at receiving terminals.

Fleet modernization: Treating voyage duration as a distribution makes the performance signal carried by vessel speed, age, and fuel-efficiency characteristics a quantifiable input to fleet renewal decisions rather than a qualitative argument.

Digital transformation: The framework outputs full duration distributions with interpretable planning percentiles, ready for downstream routing, scheduling optimization, and digital twin systems. The route-agnostic architecture scales through pooled training and shared weather features, without requiring bespoke development per corridor.

Taken together, these priorities point toward a single shift: from voyage planning as a deterministic estimate absorbed into a fixed buffer, to voyage planning as a probabilistic input that explicitly quantifies risk and feeds the broader optimization stack.

References

Adland, R., & Jia, H. (2018). The impact of weather routing on ship speed and fuel consumption. *Transportation Research Part D: Transport and Environment, 58*, 342–354. https://doi.org/10.1016/j.trd.2017.12.012

Angelopoulos, A. N., & Bates, S. (2022). *A gentle introduction to conformal prediction and distribution-free uncertainty quantification.* arXiv:2107.07511.

Faltinsen, O. M. (1990). *Sea loads on ships and offshore structures.* Cambridge University Press.

Harsha, M. M., & Mulangi, R. H. (2021). Probability distributions analysis of travel time variability for the public transit system. *International Journal of Transportation Science and Technology, 10*(4), 292–302. https://doi.org/10.1016/j.ijtst.2021.10.006

Hersbach, H., Bell, B., Berrisford, P., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society, 146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Holthuijsen, L. H. (2007). *Waves in oceanic and coastal waters.* Cambridge University Press.

IBM. (n.d.). *Monte Carlo simulation.* IBM. Retrieved May 3, 2026, from https://www.ibm.com/think/topics/monte-carlo-simulation

International Maritime Organization. (2017). *Revised guidance on weather routing* (MSC-MEPC.3/Circ.3). IMO.

Komen, G. J., Cavaleri, L., Donelan, M., Hasselmann, K., Hasselmann, S., & Janssen, P. A. E. M. (1994). *Dynamics and modelling of ocean waves.* Cambridge University Press.

Liang, M., Su, J., Liu, R. W., & Lam, J. S. L. (2024). AISClean: AIS data-driven vessel trajectory reconstruction under uncertain conditions. *Ocean Engineering, 306*, Article 117987. https://doi.org/10.1016/j.oceaneng.2024.117987

Naik, B., Tufano, P., Tupper, L. L., Ellis, S. D., & Carrese, B. (2018). Resampling methods for estimating travel time uncertainty. *Transportation Research Record, 2672*(42), 137–147. https://doi.org/10.1177/0361198118792124

Pellemans, M., Robenek, T., & Van Woensel, T. (2025). Estimating vessel arrival times in global supply chains. *Machine Learning with Applications, 19*, Article 100751. https://doi.org/10.1016/j.mlwa.2025.100751

Press, W. H., & Teukolsky, S. A. (1988). Kolmogorov-Smirnov test for two-dimensional data. *Computers in Physics, 2*(4), 74–77.

WAMDI Group. (1988). The WAM model: A third generation ocean wave prediction model. *Journal of Physical Oceanography, 18*(12), 1775–1810. https://doi.org/10.1175/1520-0485(1988)018<1775:TWMTGO>2.0.CO;2

Wu, L., Xu, Y., & Wang, F. (2020). Identifying port calls of ships by uncertain reasoning with trajectory data. *ISPRS International Journal of Geo-Information, 9*(12), Article 756. https://doi.org/10.3390/ijgi9120756

Zhang, J., Ren, X., Li, H., & Yang, Z. (2022). Incorporation of deep kernel convolution into density clustering for shipping AIS data denoising and reconstruction. *Journal of Marine Science and Engineering, 10*(9), Article 1319. https://doi.org/10.3390/jmse10091319