Blue Modeling Standard

a Benchmarking Methodology for
Ship Performance Models

Blue Modeling Standard:
a Benchmarking Methodology for Ship Performance Models

Created: 15/04/2023
Last updated: 21/12/2023

Authors: Michaël Deschoolmeester, Casimir Morobé
Contact: michael@toqua.ai, casimir@toqua.ai

Introduction

On multiple occasions we've received the question: 'How do we evaluate & compare different ship performance models?'. Ship performance models are also known as 'vessel performance models', 'speed-fuel models', or sometimes even 'consumption tables'. It is not straightforward how to measure model accuracy and compare models against one another.

The aim of this document is to provide a practical framework for addressing these questions. We at Toqua use this same framework for our own research, as can be seen in the papers we have published on the topic:

1. A Comparative Analysis on In-service Ship Monitoring Data for Modeling the Speed-Power Relation (ResearchGate, 2022)
2. Comparing the accuracy of different Ship Performance Models (HullPic, 2022)
3. How to validate the savings potential of voyage optimization?. (Toqua, 2022)

We strive to maintain a neutral and rigorous approach in this document while ensuring it is accessible to readers with limited modeling and/or coding experience. As we're trying to create an industry standard approach, we are open to any suggestions that could enhance the quality of this framework.

This document is accompanied by hands-on Python code demonstrating the methodology on a sensor dataset. This interactive use-case with example code can be accessed here.

Following the methodology, a table similar to the simplified example below will be obtained for each model. For every received data sample, the model predicts a fuel consumption. These predictions are compared to the received values, which gives an accuracy score. To make the benchmark fair and verifiable, the model should also predict the fuel consumption on an unseen 'test' dataset, whose values are known to the data owner, but hidden from the model provider.

	Received Data						**Predictions**	**Accuracy**
	Location	**Draft**	**Wave height**	…	**Speed Over Ground**	**ME Fuel**	**Predicted ME Fuel**	**ME Fuel**
Train	…	8m	1.2		12.3	20	19.5	97.5%
	…	15m	1.0		13.2	23	23.4	98.3%
	…	15m	2.1		14.1	25	26.2	95.2%
Test	…	15m	0.5		12.7	?	20.1	?
	…	10m	2.6		15.3	?	29.3	?

The Basics

1) What is a ship performance model?

Ship performance models are models that can estimate how much fuel a ship will consume at a certain speed, given the weather and operational conditions.

Example: At a speed of 12 kn, a draft of 12 m, a wave height of 2 m and a head wind of 12 kn, a ship performance model might estimate that a ship will consume 40 mt/day.

Many types of ship performance models exist: sea trial curves (often extended with correction factors as in the ISO15016/ISO19030 standards), Computational Fluid Dynamics (CFD), pure machine learning models, Toqua's Ship Kernels, etc. Appendix A provides a non-exhaustive overview of possible modeling approaches.

2) How do we know a model is accurate?

A model is only useful if it is accurate. We can measure accuracy by comparing model predictions against the 'ground truth'. Ideally, operational sensor data should be the ground truth. Accuracy is then derived by comparing model predictions against actual operational data for observed conditions. This results in an "accuracy score".

Example: For a 24h period with given weather conditions and operational conditions, the model estimates a consumption of 60 mt/day. In reality the ship that actually sailed these conditions only consumed 50 mt/day. The model thus made an error of 10 mt/day, resulting in a relative error of 20% (10/50=0.2). Or, conversely, the model has an accuracy score of 80% (1-0.2=0.8).

3) How to compare models?

The most important criteria to judge a model is the accuracy score. When comparing accuracy scores, it's important that exactly the same scope, dataset, and accuracy score are used for each model. The remainder of this document describes how to achieve this apples-to-apples comparison.

Apart from model accuracy, there are many secondary, qualitative factors that should also be considered when deciding which model best fits your specific situation. An extensive list of these qualitative criteria is provided in Appendix C: Qualitative Criteria.

4) How to challenge accuracy scores?

Among academia and solution providers, many high accuracy claims go unchallenged. However, many of these high accuracy claims would not survive some basic questions. Accuracy scores mean nothing without context.

Appendix D provides 5 key questions to challenge accuracy claims and verify if the accuracy was derived in a correct and meaningful way for the application at hand.

Methodology: how to calculate model accuracy

First, we establish a benchmark dataset
This dataset will provide the ‘ground truth’ of measured values that the models should try to estimate. It should be free of errors and constrained to sea-going conditions.

Next, we make predictions and calculate quantitative metrics
To measure how accurate the models are it is important to have meaningful, robust and quantifiable metrics. It is important that the metrics for the evaluated models are calculated for the exact same dataset.

Finally, we benchmark the results
We compare the obtained quantitative metrics against each other to determine the most accurate model.

1. Establishing a benchmark dataset

The goal is to establish a dataset upon which all models can be compared. Most importantly, **the exact same dataset should be used for all models.**

1) **Choose a data source: noon reports or sensor data**
Sensor data is preferred as it has the highest frequency and is most accurate (assuming data quality is properly monitored.). All variables required by the models should be available.

2) **Choose an evaluation period**
Constrain the dataset to a period during which ship performance is assumed to be stable. A longer period will lead to better estimates of accuracy scores. It should be free from long idling periods, cleaning events, dry dockings or ESD-installations. Fouling should not be present, or remain at a constant level.

3) **Filter outliers**
Remove any erroneous measurements from the dataset. It’s important to only filter data which is known to be incorrect. Data for conditions that are simply more challenging to model (e.g. bad weather, uncommon drafts, uncommon speeds) should not be removed. Many outlier detection algorithms exist, a discussion of which is out of scope.

4) **Filter unaccounted for external effects**
Remove any measurements for which there are effects at play that influence ship performance but are not taken into account by the models. For example: shallow waters,

icing and large rudder angles. We recommend the following set of filters as starting point:

- Sea Depth > 3 * mean draft
- Rudder Angle < |2 degrees|
- Sea Temperature > 0 degrees

5) **Constrain to sea-going conditions**
Remove any measurements during which the ship is not in normal operating conditions at sea, e.g. where it is manoeuvering or anchoring. We recommend the following set of ship-dependent filters that already cover most cases:
   - Speed > ... kn
   - RPM > …
   - Power > … kW
   - Fuel Consumption > … mt/day

What's left should ideally be a dataset with **at least 2 months of data over at least 2 voyages, covering both laden and ballast conditions**.

Important: When calculating accuracy metrics for a model that used operational data to learn from, it is important that the chosen data set has not also been used to train the model. For more information we advise reading the concept of Training, Validation, and Test data sets.

2. Determine quantitative metrics

To quantify how well a model performs, we focus on quantitative metrics that are comparable across models and across ships. We consider the ones below most important. The exact formulas of these metrics are provided in Appendix B.

Metric	Description
MAPE	**Mean Absolute Percentage Error** The average absolute error per prediction, relative to the measured value. Example frequencies could be 1 min, 5 min, 1 hour, etc. **Example** Every hour for 3 hours, a value of 50mt/day is measured as true value, and thus true consumption by the vessel. If the model predicts 40mt/day for the first hour, then 55mt/day for the second hour, and 50mt/day for the last hour, then the corresponding Percentage Errors are: -10/50 = -20%, 5/50 = 10% 0/50 = 0% The resulting Mean Absolute Percentage Error is (|-20%|+|10%|+|0%|)/3=10%
BPE	**Bias Percentage Error**

	The error averaged over all the data. This gives an indication of how much the model over or underestimates on average (i.e. the bias). **Example** The total consumption over a 3 month period was measured at 3000 mt. If for the same period, the model predicted the aggregate consumption would be 3300mt, that’s an overestimation of 300mt. That equals a BPE of +10%.
DPE	**Daily Percentage Error** Similar to MAPE, but averages errors over a 24h period (days). As overpredictions and underpredictions can partially cancel each other out over time, this error over 24h will be lower than the error at a 1hour or 5 min frequency. **Example** The true consumption for a ship was a constant 50mt/day throughout the day. The model predicts a consumption of 55mt/day for the first 12 hours of the day, and 45mt/day for the last 12 hours of the day. The aggregated value over 24h is thus also 50 tons. The DPE will be 0%. Note that if we had not calculated the DPE (MAPE per 24h), but the MAPE per 1h, that the MAPE per 1h would be 10%, while the DPE is 0%.
VE	**Voyage Error** The same as BPE, but calculated per voyage. This gives an indication of how much the model over or underestimates on a voyage. **Example** A 20-day voyage actually consumed 900mt. The model predicted this voyage would consume 810mt. The VE would be (810-900)/900 = -10%

We’d like to stress that quantitative metrics are just one part of the full picture. A non-exhaustive set of other, qualitative criteria may be found in Appendix C.

3. Benchmarking models

Provided below are the results of a **fictional (!)** benchmarking study of three models. These three models have been evaluated using our framework outlined above.

Quantitative Comparison

	Sea trial + ISO corrections	**Model Z**	**Ship Kernel**
MAPE (5 min)	18%	13.6%	6.9%
DPE	15.3%	11.7%	2.4%
VE	10.8%	6.9%	1.0%
BPE	8.4%	5.1%	-0.4%

Qualitative Comparison

As accuracy metrics are just one factor in comparing models, we've determined a set of qualitative criteria to judge models upon. These can be found in Appendix C. We have assessed these criteria for the Sea Trial + ISO corrections (applying ISO 15016 & ISO 19030) and compared against Ship Kernels. Those results can be found in the respective subsections of Appendix C. The subsection on Model Z is left intentionally blank such that readers may apply it to their own model.

Example conclusion of fictional benchmarking study

When looking at the quantitative comparison based on accuracy scores, the Toqua Ship Kernel clearly comes out on top, while the Sea Trial + ISO corrections model is the worst performer. On a voyage level, the Ship Kernels have an average error of 1.0% while the Sea Trial + ISO corrections has a much higher average error of 10.8%. For the use-case of routing, model accuracy is by far the most deciding factor (Vanackere, Guatama, and Morobé, 2022). So if the purpose is to create more accurate ship performance models to increase fuel savings of routing, then Ship Kernels would be the obvious choice.

However, the Sea Trial-based approach has the qualitative benefits of being easy to interpret and that it does not require any sensor data. Depending on the use case these qualitative benefits might justify the simpler and less accurate approach.

Conclusion

The goal of this document is to standardize industry-wide discussions regarding ship performance models and their accuracy. This with the purpose of improving collaboration and opening discussions on the topic among industry peers and academia.

A standard only becomes industry-standard over time as adoption and acceptance increases among different industry players. As such, this standard is very open to feedback and suggestions from the industry.

If you would like to underwrite this standard or contribute, feel free to get in touch via bluemodelingstandard@toqua.ai.

Appendices

Appendix A: Different Modeling Approaches

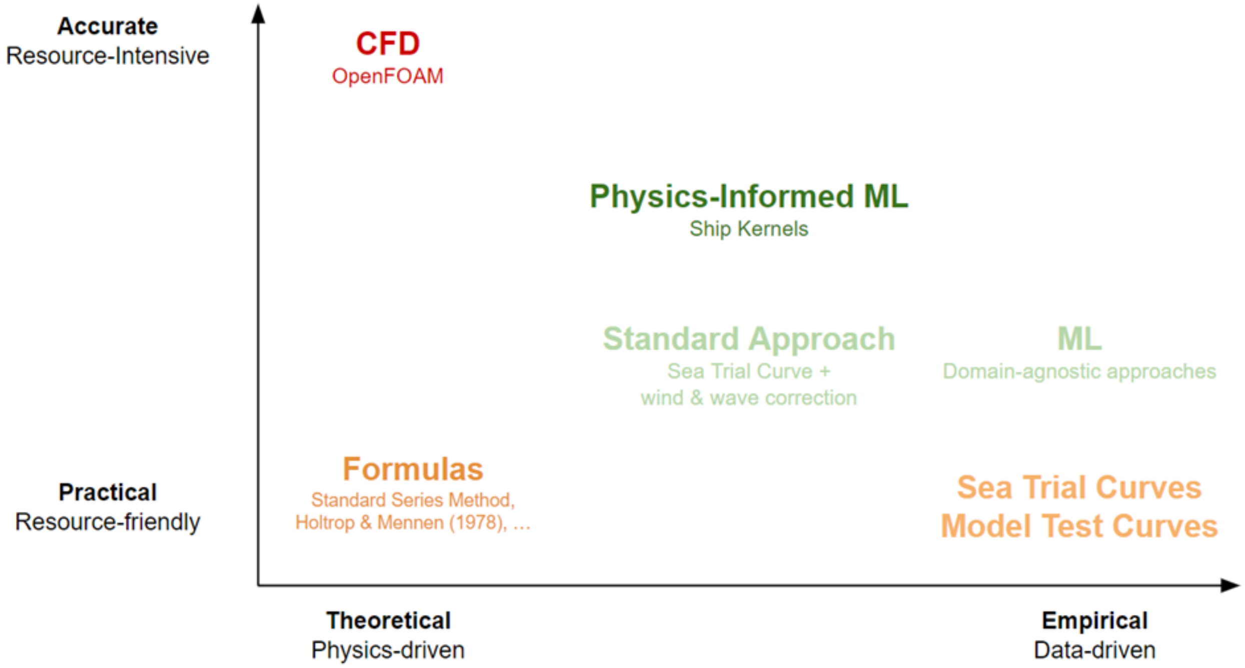


Figure 1: Different Modeling Approaches. (Colle and Morobé, 2022)

Appendix B: Quantitative metrics

In the formulas below, y_i is the true value as it is in the dataset, while $\hat{y}_i$ is the predicted value by the model.

Metric	Description
MAPE	**Mean Absolute Percentage Error** $$\mathrm{MAPE} = \frac{100}{N} \sum_{i=1}^{N} \left\| \frac{\hat{y}_i - y_i}{y_i} \right\|$$
DPE	**Daily Percentage Error** $$\mathrm{DPE} = \frac{100}{D} \sum_{d}^{D} \left\| \frac{\sum_i^d \hat{y}_i - y_i}{\sum_j^d y_j} \right\|$$ D denotes the number of days and d the measurements in a day
BPE	**Bias Percentage Error** $$\mathrm{BPE} = \frac{\sum_i^N \hat{y}_i - \sum_j^N y_j}{\sum_k^N y_k}$$
VE	**Voyage Error** $$\mathrm{VE} = \frac{1}{V} \sum_{v}^{V} \left\| \frac{100}{v} \sum_{i}^{v} \frac{\hat{y}_i - y_i}{y_i} \right\|$$ V denotes the number of voyages and v the measurements in a voyage

Appendix C: Qualitative Criteria

Criteria	**Description**
Data requirements	What data is required to create the model? How difficult/costly is it to obtain this data? How much training data is required?
Dimensionality	Which input parameters does the model take into account? This helps understand what performance-influencing variables might not be accounted for by the model. **Example** draft, trim, (significant) wave height, wave angle, wave period, wind speed, wind angle, water depth, sea salinity, sea temperature, etc.
Prediction Time	How long does it take to execute a prediction? How does this scale with increasing sample size?
Input Flexibility	Starting from what variables can the model make a prediction? *Traditionally models predict from Speed Over Ground (SOG) towards Main Engine Fuel Consumption. But depending on the use case it might be useful to start predicting from RPM towards speed, or from Fuel towards SOG.* **Example inputs** SOG, STW, RPM, main engine power, main engine fuel consumption
Output Flexibility	Which output parameters can the model predict? **Example outputs** STW, SOG, RPM, main engine power, main engine fuel consumption, boiler consumption, …
Physical Robustness	How guaranteed is it that the model will respect physical truths? How is physical correctness guaranteed?
Uncertainty	Is there a mechanism to know the uncertainty inherent to each prediction? How interpretable is this mechanism?
Explainability	How interpretable are the model's predictions? Can the way in which a model arrived at a prediction be explained adequately?
Dynamicity	Can the model dynamically adapt to the changing performance of a ship over time? How frequently does it update?

Applicability	In which contexts can the model (not) be used?

Sea Trial + ISO corrections

Data requirements	Requires sea trial documents. Requires various ship-specific theoretical parameters Depending on the method also requires NR or sensor data.
Dimensionality	Depends on implementation details of ISO15016. Could for example be draft (laden or ballast), wind speed and wave height.
Prediction Time	Near instantaneous
Input Flexibility	STW
Output Flexibility	Power
Physical Robustness	Very robust. The curves are created from actual sailing conditions. The corrections are fitted from physical formulas. As such, the model is as good as guaranteed to be physically robust, but not per se accurate.
Uncertainty	No uncertainty mechanism present. ISO 15016 & ISO 19030 do not contain a section on uncertainty of predictions.
Explainability	Very explainable. As the curves are simple, and the corrections are formulas you can deduce the output immediately from all inputs.
Adaptability	Can adapt over time using additional corrections. Typically happens infrequently, as it requires human effort and attentiveness.
Applicability	Determining ship over or underperformance, speed-fuel tables, Charter Party performance, sometimes also used in routing algorithms.

Ship Kernels

Data requirements	Requires the logging of sensor data. Depending on the case, at most a few months of sensor data is required, in the best case only a few days.
Dimensionality	Draft, trim, wave height, wave direction, wave period, wind speed, wind direction, current speed, current direction, sea surface temperature, sea surface salinity, calorific value of the fuel, etc. Depends on data availability and ship type.

Prediction Time	Near instantaneous for a single prediction (~200ms). Scales sub-linearly: less than 3s for 10,000 predictions.
Input Flexibility	SOG, STW, RPM, Power, Fuel consumption
Output Flexibility	Starting from any of the input variables above, all following outputs can be predicted: SOG, STW, RPM, Power, Fuel consumption, CO_2 emissions
Physical	Plausible. Ship kernels are based on physics-informed machine learning. A hybrid of naval architecture and physical rules are embedded in the learning architecture of the model, guaranteeing physically correct predictions. An additional framework of validation steps ensures this is true for every model.
Uncertainty	Possible.
Explainability	Depends on predicted parameters
Adaptability	Yes. Automatically adapts over time to the latest actual performance of the ship. Updates can happen on a daily basis, depending on how often new data arrives. Models always keep track of historical performance, allowing for historical simulations.
Applicability	Determining ship over or underperformance, Charter Party performance, routing, speed optimization, speed-fuel tables, simulations of any kind, etc.

Model Z (template)

Data requirements	
Dimensionality	
Prediction time	
Input Flexibility	
Output Flexibility	
Physical Robustness	
Uncertainty	
Explainability	
Adaptability	
Applicability	

Appendix D: Questions to challenge accuracy claims

1. **What data is considered the ground truth?**
   Typical answers include:
   -Sea Trial Curve
   -Computational Fluid Dynamics (CFD)
   -Operational Noon Report Data
   -Operational Sensor Data

   The right answer depends on the case the models will be used for.
   For operational applications such as routing optimization and hull performance monitoring, the goal is to analyze performance of the model in many different operational conditions. In that case, Sea Trial Curves and CFD are not suited, as they only represent a very limited amount of conditions. For those cases only operational noon report data and sensor data over longer periods of time are valid to serve as ‘ground truth’. Assuming data quality is properly monitored, sensor data is preferred over noon report data, due to higher frequency and accuracy.

   Lastly, what filtering has been applied to the data considered to be the ‘ground truth’? For example, some models filter out all data above 4BFT while others don’t and attempt to model all weather conditions. Heavy filtering will lead to better accuracy scores. Ideally, as little filtering as possible happens, so the models are useful for all possible operational conditions the vessel might experience..

2. **What model validation technique is used?**
   Model validation is the set of processes intended to verify that models are performing as expected. This question is mostly relevant when analyzing claims for operational model accuracy, especially for techniques like machine learning or other modeling approaches that update/calibrate the models based on actual operational data.

   Typical answers include:
   -train set = test set (overfitting/leakage)
   -train/test split
   -train/validation/test split
   -cross-validation
   -customer-led blind test

   There are many valid approaches to model validation, but the same validation technique needs to be used over different models to ensure comparability.

   The only truly incorrect answer is the first one where the train set is equal to the test set. This is called ‘overfitting’ or ‘leakage’ and is one of the most common mistakes when analyzing model accuracy. It's crucial to watch out for this. In essence, it occurs when the data the model is tested for, has also been used to train the model in the first place.

A classic example in shipping is when a speed-fuel model is being retrained/recalibrated throughout the voyage based on operational data, and then accuracy is reported at the end of the voyage using that recalibrated model. This will lead to misleadingly high accuracy scores that the model can not achieve in practice to predict voyage performance before the voyage starts.

To avoid any mistakes, ensure fairness, and ensure comparability over different modeling approaches, this standard recommends companies to choose for a '**customer-led blind test**'. This means companies share X months of data to different modelling providers to train the model, while keeping the Y consecutive months to themselves to test the models. This removes the risk for overfitting completely, as the test data has never been in possession of the modeling provider. Important is that X and Y months together make up a period of relatively 'stable performance' without any major cleaning events or accumulation of idling that could cause a meaningful shift in vessel performance.

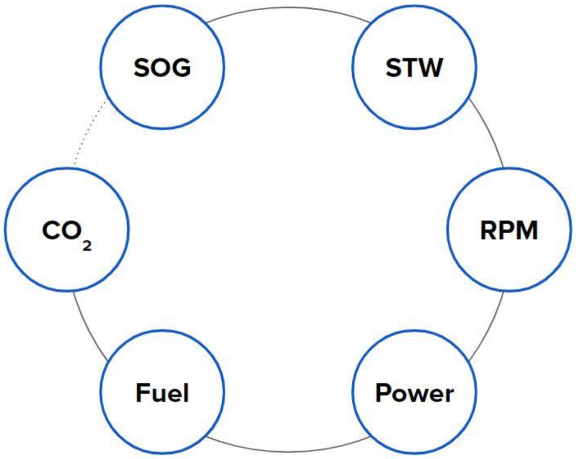
If customers are unable to make the required model predictions on their side, an alternative approach is to share only the input variables for the Y months of test data with the modeling provider, while dropping the dependent variables. Then the modeling provider makes the predictions on their side, and returns the document including the predictions, so the company can compare these against the actual values on their side.

For example, if the model aims to predict Fuel Consumption, starting from Speed-Over-Ground, then Y months of data could be shared with the modeling provider, only including the Speed-Over-Ground, location, timestamp, draft, weather, etc. Variables such as RPM, Power, Fuel Consumption, etc. should not be included, but should be predicted by the modeling provider.

3. **What relationship is modeled?**

   It is common to model the fuel, given the speed as an input, but there are many more possibilities.

   Typical answers include:
   Model fuel, starting from SOG.
   Model SOG, starting from fuel.
   Model power, starting from RPM.
   Model power, starting from RPM and SOG.
   etc.



To ensure comparability, the same relation should be modeled in the same direction. Modeling SOG starting from Fuel will lead to much higher % accuracy scores than modeling Fuel starting from SOG, for example, due to the exponential nature of the relation.

4. **What time horizon is used for the accuracy?**
   Typical answers include:
   -Point prediction (1s-15min depending on data frequency)
   -Day
   -Week
   -Month
   -Voyage
   -Whole Period

   The right answer depends on the case the models will be used for.
   For example, if the goal is to predict the fuel consumption of a voyage, before the voyage has started, then accuracy should be analyzed on a voyage level.
   In general, the longer the period, the better the accuracy.

5. **What accuracy metric is used?**
   See Appendix B: Quantitative metrics.

   Apart from the ones listed in Appendix B, typical answers include:
   -R^2 (determination coefficient)
   -MAE
   -MSE
   -RMSE

   This standard recommends relative metrics, such as MAPE, BPE, and VE, expressed in percentages. This allows for better comparability over different vessel types and is easier to understand from a user perspective.

Template: Questions to challenge accuracy claims

Modeling Scope

What data is considered the ground truth?
e.g. Sea Trial data/Noon reports/Sensor data/etc.
What model validation technique is used?
e.g. A train/test split is performed in the following way…
What relationship is modeled?
e.g. SOG→Fuel / Power→STW / RPM→Power, etc.
What time horizon is used for the accuracy?
e.g. Point prediction/Day/Voyage/Whole period/etc.
What accuracy metric is used?
e.g. MAE/MAPE/MBE/etc.