Fuel Model Benchmarking Report: Provider X

ZeroNorth - September 6, 2021

Scorecard

Weather	Cargo / Speed	Ballast Low	Ballast Medium	Ballast High	Laden Low	Laden Medium	Laden High
Bad	Accuracy %	67.3	81.3	85.4	70.7	85.6	89.5
	Std	18.4	13.0	11.7	15.8	10.6	8.7
	Bias %	-29.4	-11.7	-2.5	-27.6	-9.5	-3.9
	# Noons	7805	27280	6286	6815	31639	5566
Good	Accuracy %	66.6	80.8	85.3	64.1	84.1	88.2
	Std	20.0	13.5	11.6	18.5	11.9	9.6
	Bias %	-28.3	-11.8	-4.6	-30.9	-11.3	-6.8
	# Noons	5184	22324	6469	2294	21645	5638

Table 1: The unweighted accuracy on noon reports by various categories. Good and bad weather is defined as a value below and above 4 on the Beaufort scale respectively. Low, medium, and high speed is defined as the vessel moving below 10, between 10 and 13, and above 13 knots.

The scorecard gives a summarized overview of the performance of the fuel table provider in various categories. These categories have been chosen to give the full picture of a provider's performance in rare and common conditions. Further details about the best and worst performance in categories with finer granularity is given in Table 2 and Table 3.

Introduction to Provider X as a fuel table provider

Provider X's fuel models are partly based on machine learning trained on noon reports but with naval architecture as the backbone of the fundamental physical relationships. The following parameters are used by Provider X:

- Speed over ground [knots]
- Draught [m]
- Wind speed [m/s] / Wind Direction
- Latest drydock date

ZeroNorth integrates with Provider X through an API.

Performance Analysis

In this section, the performance analysis for the fuel table provider is introduced. The analysis was only applied to vessels where atleast 50 noon reports were available. The reason for this only showing the analysis on vessels with atleast can be seen in Figure 1, where vessel level weighted accuracy is plotted against the number of available noon reports for the vessel. The figure shows that when the number of noon reports increases, the variance in the accuracy decreases, which in turn means a more reliable analysis.

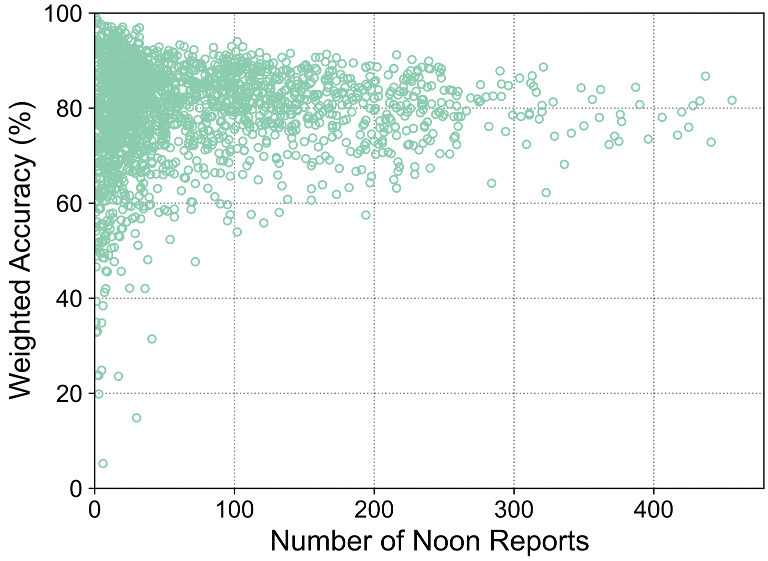
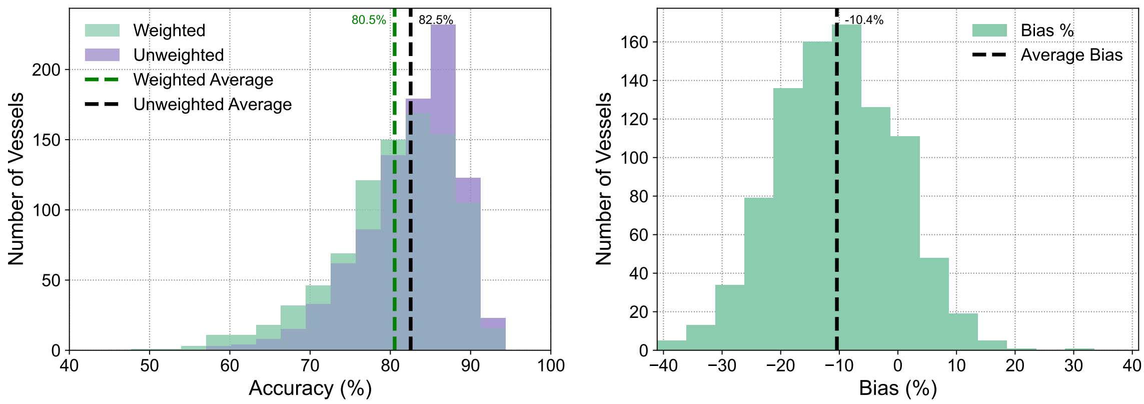


Figure 1: Scatter plot showing the weighted accuracy % on a vessel level, plotted versus the number of noon reports per vessel. As the number of noon reports increases the variance in the weighted accuracy % across all vessels is reduced.

In Figure 2 the per vessel histogram of the accuracy and bias is shown. In Figure 2a, the weighted and unweighted accuracy distributions are shown with their corresponding mean values shown as vertical dashed lines. In Figure 2b, the bias distribution is shown with its corresponding mean value as the dashed line. A negative bias indicates that the fuel models on average predicts a lower consumption than reported in the noon reports of each vessel. For more details on the metrics, the reader is referred to the technical descriptions above.



(a) Accuracy distribution

(b) Bias distribution

Figure 2: Distributions of accuracy and bias per vessel, for vessels with a minimum of 50 noon reports.

In Figure 3, a boxplot of the standardised mean absolute error on a noon report level is shown, split by groupings of speed and beaufort levels. This plot can be used to assess the general performance of the fuel table provider in different vessel and sea states.

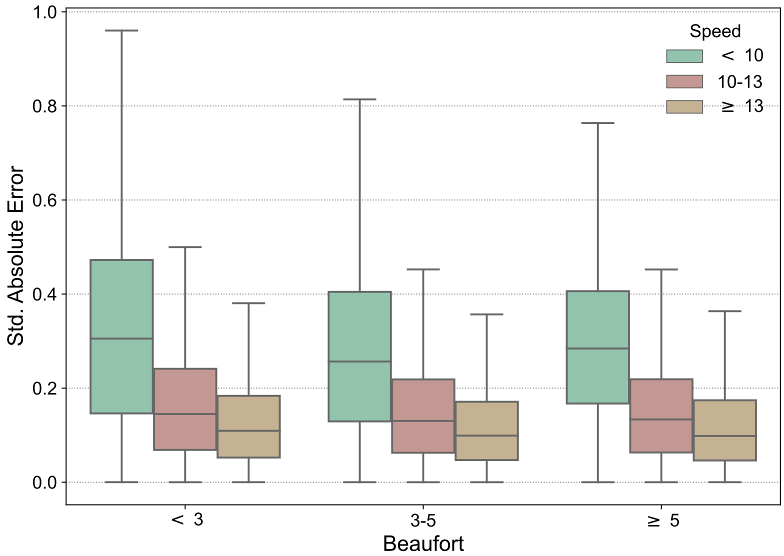


Figure 3: Boxplot showing the standardised mean absolute error on a noon report level across all vessels, split by groupings of speed and beaufort.

In Table 2 and Table 3 an overview of the unweighted accuracy and bias is given on a per bin level, for the bottom and top performing bins across all vessels. These tables should be used to give a deep dive into how each fuel table provider performs in various rare and common conditions, when benchmarking their fuel models against noon report consumptions.

DWT	Speed	Beaufort	Draught	Accuracy %	Bias %	# Noons
< 50000	< 10	< 3	9-12	60.87	-33.45	229
	< 10	< 3	< 9	61.70	-33.22	1148
	< 10	≥ 5	< 9	63.07	-34.40	2252
	< 10	3-5	< 9	64.68	-31.15	3107
	< 10	3-5	9-12	65.31	-30.90	860
	< 10	≥ 5	9-12	68.21	-29.87	998
	10-13	< 3	< 9	79.38	-14.19	3960
	10-13	3-5	< 9	80.57	-13.28	14385
	10-13	≥ 5	< 9	80.89	-12.35	4915
	10-13	< 3	9-12	83.46	-9.39	1838
50000-150000	< 10	< 3	12-16	62.42	-32.01	205
	< 10	< 3	9-12	62.96	-29.99	721
	< 10	< 3	< 9	63.03	-30.01	172
	< 10	≥ 5	< 9	63.33	-35.28	423
	< 10	3-5	< 9	66.54	-27.41	562
	< 10	3-5	9-12	69.13	-26.47	2766
	< 10	≥ 5	9-12	69.44	-29.10	2926
	< 10	3-5	12-16	69.91	-26.33	788
	< 10	≥ 5	12-16	74.91	-24.00	1081
	10-13	< 3	< 9	78.27	-16.97	710
≥ 150000	< 10	< 3	12-16	63.28	-32.32	298
	< 10	3-5	≥ 16	65.13	-33.21	76
	< 10	≥ 5	≥ 16	65.80	-31.62	52
	< 10	≥ 5	12-16	67.75	-30.31	712
	< 10	≥ 5	< 9	69.18	-28.97	100
	< 10	≥ 5	9-12	70.99	-25.43	435
	< 10	3-5	12-16	71.76	-24.75	937
	< 10	< 3	9-12	73.64	-22.59	255
	< 10	3-5	< 9	74.41	-23.79	160
	< 10	3-5	9-12	76.23	-18.92	748

Table 2: Bottom 10 bins for each vessel category, sorted ascending by their unweighted accuracy %, defined as the average across all noon reports within the bin per vessel category. Bins where less than 50 noon reports were available were excluded from the overview.

DWT	Speed	Beaufort	Draught	Accuracy %	Bias %	# Noons
< 50000	≥ 13	3-5	9-12	87.66	-3.81	1839
	≥ 13	≥ 5	9-12	87.39	-1.13	431
	≥ 13	3-5	< 9	86.10	-3.95	3098
	≥ 13	< 3	< 9	85.69	-5.19	847
	≥ 13	< 3	9-12	85.58	-5.10	512
	≥ 13	≥ 5	< 9	85.08	1.36	711
	10-13	3-5	9-12	84.38	-9.07	8050
	10-13	≥ 5	9-12	84.17	-8.87	2977
	10-13	3-5	12-16	84.02	-9.39	74
	10-13	< 3	9-12	83.46	-9.39	1838
50000-150000	≥ 13	≥ 5	12-16	90.61	2.04	307
	≥ 13	3-5	12-16	89.82	-0.60	1184
	≥ 13	< 3	12-16	88.33	-3.98	370
	≥ 13	3-5	9-12	88.18	-3.24	6355
	≥ 13	≥ 5	9-12	87.94	-0.90	1531
	10-13	≥ 5	12-16	87.38	-5.95	3257
	≥ 13	≥ 5	< 9	86.73	-6.78	357
	≥ 13	< 3	9-12	86.35	-6.05	1825
	10-13	3-5	12-16	86.13	-8.41	7163
	≥ 13	3-5	< 9	85.69	-10.47	1270
≥ 150000	≥ 13	3-5	≥ 16	88.42	-6.85	135
	10-13	3-5	< 9	86.64	-9.99	281
	≥ 13	≥ 5	12-16	86.47	-0.62	313
	≥ 13	3-5	12-16	85.43	-5.76	1212
	≥ 13	≥ 5	9-12	85.15	-4.37	149
	10-13	< 3	< 9	84.86	-11.20	69
	≥ 13	3-5	9-12	83.93	-10.22	566
	10-13	≥ 5	≥ 16	83.84	-12.89	279
	≥ 13	< 3	9-12	83.61	-8.69	150
	≥ 13	< 3	12-16	83.36	-7.80	276

Table 3: Top 10 bins for each vessel category, sorted descending by their unweighted accuracy %, defined as the average across all noon reports within the bin per vessel category. Bins where less than 50 noon reports were available were excluded from the overview.

How ZeroNorth measures fuel model accuracy

The goal of this report is to provide a deep-dive into the performance of one of the various fuel model providers that ZeroNorth collaborates and integrates with. ZeroNorth provides a voyage optimisation tool used to optimise voyage TCEs and to do so, our tools and algorithms are built with the assumption that the underlying fuel models are reasonable accurate, reliable and reflects real world scenarios.

The optimisation algorithm simulates many different ETAs for a voyage leg and based on the estimated consumption from the fuel model and combined with the market conditions, the algorithm will find an optimal solution. Hence for the algorithm to find solutions that reflects real world scenarios, it is evident that the fuel model must be good at modelling the vessel at different speeds, weather and vessel conditions.

Given that a fuel model needs to be able to model the vessel well in all these conditions, ZeroNorth has decided to develop a weighted accuracy metric to evaluate the performance of a fuel model, which is introduced below.

Technical Approach

In this report we introduce two metrics for evaluating the performance of a fuel model. The first metric is defined in (1) and is a normalised mean absolute error presented as an accuracy. Here, N denotes the number of samples in the group, $y_{p,i}$ and $y_{t,i}$ denotes the predicted and reported fuel consumption for the i'th sample, respectively, and w_i is a weight given to sample i. The metric is unweighted if $w_i = 1$ for all samples.

$$100\% \cdot \left(1 - \frac{\frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} w_i \left| y_{p,i} - y_{t,i} \right|}{\frac{1}{N} \sum_{i=1}^{N} y_{t,i}} \right) \qquad (1)$$

A specific weight w_i is applied to each sample i in (1). As described, it is paramount that a fuel table performs well across the full parameter space when ZeroNorth's optimisation algorithm is searching for an optimal voyage plan. To reflect this in the performance on the weighted accuracy metric, the weighting scheme applied tries to reflect this by splitting the parameter space into several bins, each reflecting different vessel and sea states. The parameter space is separated into the combination of bins as listed below, totalling 108 bins across the parameter space. Then, using the full dataset of all noon reports across ZeroNorth's customers, a n-dimensional histogram is calculated, counting the number of noon reports within each bin.

In the following performance analysis, the bins used are as listed below:

- Bins of speeds [kn]: < 10, 10-13, and ≥ 13
- Bins of beaufort: < 3, 3-5, and ≥ 5
- Bins of draught: < 9, 9-12, 12-16, and ≥ 16
- Bins based on vessel DWT: < 50000, 50000-150000, and ≥ 150000

It should be clear to the reader that because the combination of bins reflects both common and rare scenarios of vessel and sea states, it is natural to expect a non-uniform distribution across the the parameter space.

Next, we introduce the weighting scheme. The weighting scheme is developed with an emphasis on transparency and such that it is data agnostic, whilst ensuring that rare vessel and sea states are weighted higher than common conditions. The proposed weighting scheme is as seen below in (2).

$$w_i = W_{max}^{\frac{C_{max} - C_b}{C_{max} - C_{min}}} \qquad (2)$$

Here, W_{max} is the maximum weight a bin can have. Here with have choosen to set $W_{max} = 10$, such that the most rare bin is weighted ten times as much as the most common bin. The minimum and maximum counts seen in all bins are denoted C_{min} and C_{max}, respectively such that $0 \leq C_{min} \leq$

$C_b \leq C_{max}$. The weights decrease exponentially as C_b increases and is bounded to 1 when $C_b = C_{max}$. Likewise, w_i bounded to W_{max} when $C_b = C_{min}$.

Besides measuring the accuracy of the fuel model, it is also of interest to look at another metric. In (3), the metric called bias is shown and it is an unweighted estimate. The bias basically shows whether the fuel model is consistently over- or underestimating the noon report consumption. This happens if the bias is non-zero. There can be multiple reasons if there is a consistent bias: for example a systematic reporting error in the reports or a fuel model that doesn't capture the effect of fouling on the vessel. Further, a fuel model might have a bias of zero, whilst still having a low accuracy. However, if the model has a large bias, then the model will likely also have a low accuracy.

$$100\% \cdot \left(\frac{\sum_{i=1}^{N} (y_{p,i} - y_{t,i})}{\sum_{i=1}^{N} y_{t,i}} \right) \tag{3}$$

Using noon reports for benchmarking fuel models

This section describes how noon reports can be processed such that they can be used for measuring accuracy and bias. Additionally, we present a case study where noon reports are compared to a corresponding set of sensor data, recorded on the same vessels across a large period of time.

Noon reports cover a certain period of time (although typically 24 hours) and contains information about the consumption used and the condition of the vessel as well as a number of weather- and sea inputs. In order to decrease the uncertainty of the analysis, we use external AIS and weather data to find the actual speed and weather experienced during the noon report period. Specifically, we break the noon report period into segments and estimate the consumptions on each which are then aggregated to give a total estimate for the whole noon report period. Thus, it is possible to use (1) and (3) to measure accuracy and bias, respectively. Hence, only the reported consumption and draught from the noon report are used as inputs to the fuel model, whereas speed and weather comes from external data sources. This approach allows for a fair estimate across many vessels and conditions.

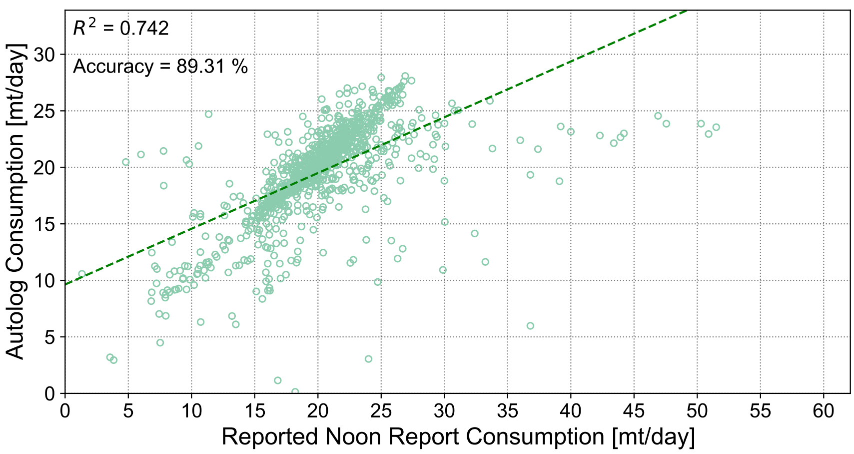


Figure 4: A case study of reported consumptions from noon reports compared to corresponding sensor data automatically collected aboard five tanker vessels, collected across a large time period. The R^2 measure is derived from the best linear fit and the accuracy is calculated as the unweighted accuracy as given in (1).

When comparing fuel models with noon report data, one should realise that the reported consumptions within the noon reports might be affected by different types of noise. One source of noise which is usually present in the data is stemming from human error, meaning that the reported consumption might be inaccurate or wrong. However, given a large enough dataset, we believe a fairly good representation of the fuel model accuracy can be achieved. Because of this noise, which might vary across datasets and vessels, it is reasonable to expect that the accuracy of a fuel model, when benchmarked against noon report data, is upper bounded at some level below 100 % on average. In Figure 4, the result of a case study on how noon report consumptions match consumptions as recorded on sensor data, for 5 tanker vessels across a larger time period. When computing the unweighted accuracy between two datasets of consumption, then this is 89 % percent. This needs to be kept in mind when assessing the fuel model accuracy towards the noon reports and in the case of these five vessels, the 89 % unweighted accuracy should be considered an upper bound on the accuracy of a fuel model when benchmarking it against noon report data. Assuming this case study is representative, one cannot expect to have a fuel model with an accuracy above 89 % when measured against a large set of noon reports. However, there might be significant variations between vessels and their report quality and hence the upper bound might be different for different vessels.