Saving AI from its own hype: Getting real about the benefits and challenges of machine learning for ship performance modelling aimed at operational optimizations

Camille Colle, Toqua, Gent/Belgium, camille@toqua.ai
Casimir Morobé, Toqua, Gent/Belgium, casimir@toqua.ai

Abstract
This paper compares different approaches for ship performance modelling, with the goal of finding the modelling technique best suited for operational optimizations. Extra emphasis is placed on the potential and challenges of data-driven methods such as machine learning. The added value of using data driven methods based on sensor data compared to noon reports is quantified. Next to industry-standard approaches, a new approach based on physics-informed machine learning called 'ship kernels' is proposed. Ship kernels are shown to outperform the other approaches considered here in short-term accuracy. This makes them an ideal building block for operational optimizations (such as routing and speed optimization) that require predictions for a broad range of conditions. The ship kernels are shown to have excellent long term accuracy compared to other approaches, making them a valuable tool for performance monitoring use-cases such as maintenance planning related to hull & propeller performance. This paper concludes with general remarks and warnings on the challenges of operationalizing machine learning.

1. Introduction

The terms AI and ML have been overused and misused in the last years to the point where they have lost any meaning to most people. According to [1] 40% of AI-startups don't use AI. This fact shows how many companies want to become part of the trend. Nevertheless, using Machine Learning should not be a goal by itself. It is a means to an end. Actually, a general rule of machine learning is to start without machine learning [2]. Machine Learning requires high-quality data, robust data pipelines and costly engineers to maintain these systems in production. For most problems a simple heuristic or approximation will give comparable results for a fraction of the costs and effort. It is only in very few cases that these few extra percentages of accuracy due to Machine Learning merit the effort. This paper argues that Ship Performance Modelling for operational optimizations (routing, maintenance planning/fouling detection and speed optimization) is such a case, if sensor data is available.

2. A qualitative analysis of ship performance modelling approaches for operational optimizations

The most common modelling approaches used in the industry today are compared in Figure 1. The solution presented in this paper, developed by Toqua, is denoted "Ship Kernels". In general there is a tradeoff between accuracy and implementation cost. Below this tradeoff is discussed in detail for the different approaches.

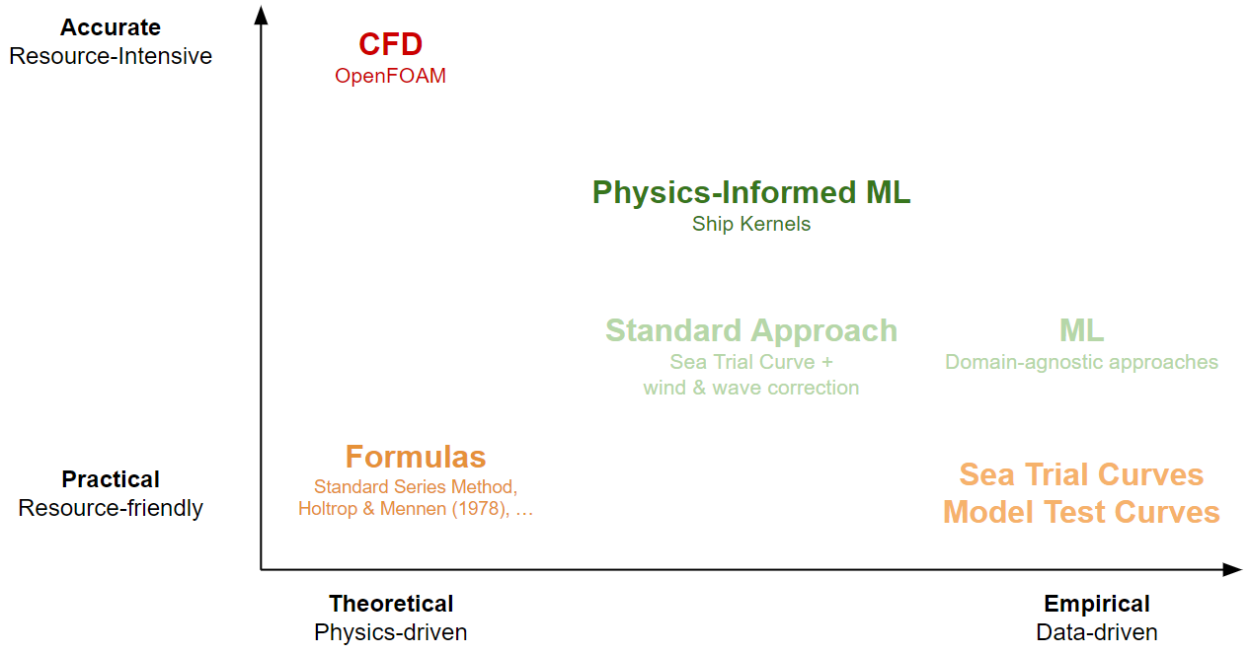

Figure 1: A pictorial overview of the different approaches in current vessel performance modelling solutions.

Computational Fluid Dynamics (CFD) are highly accurate and generate interpretable results, with the drawback of requiring very niche expertise, long computation times and high costs. As a result, CFD are considered infeasible for operational optimizations where a high number of predictions have to be made for a broad range of weather conditions in a limited timespan.

Sea trial curves are highly practical and easy to use but have limited accuracy and flexibility. They can not account for changing factors such as weather conditions, which have a large influence on the vessel's performance.

Formulas based on analytical expressions like the Standard Series Method [3] and the approximate power prediction method of Holtrop and Mennen [4] are highly transparent with low computational cost compared to CFD calculations. These expressions often require a large number of empirical coefficients not readily available for a given vessel. As such simplifications have to be made that reduce accuracy. Even though highly practical and low-cost, both sea trial curves and formulas as stand-alone solutions are considered rather inaccurate due to oversimplifications and data limitations.

Domain-agnostic ML approaches - under the motto 'chuck in some data and see what happens' - can generate fairly accurate results under the prerequisite high-quality data is available and predictions are made for operating conditions similar to training conditions. This approach may generate unreliable and unrealistic results for operational optimizations that consider weather & operating conditions not (frequently) observed in the data. This can lead to incorrect and costly business decisions.

A standard approach frequently used in the industry combines sea trial curves with theoretical formulas such as ISO-15016 [5] or Kreitner's formula [6] that corrects the sea trial curves for the added resistance

due to wind and waves. These combinations still have a limited complexity, low computational cost and can improve the accuracy compared to uncorrected sea trial curves significantly. As a result, variants of this combination are often the default ship performance model used for operational optimizations in shipping today.

Finally, this paper presents a new approach for ship performance modelling that combines the benefits of sensor data availability and traditional theoretical insights based on physics and naval engineering. This is the domain of Physics-Informed Machine Learning [7], a new and quickly developing domain aiming to combine the benefits of data-driven and physics-driven approaches. Toqua has developed Physics-Informed Machine Learning models for ship performance, called 'Ship Kernels'. Given data of sufficient quality, these ship kernels can outperform sea trial curves, formulas, domain-agnostic ML and correction-based approaches, while still being highly flexible, low-cost and reliable. We argue that ship kernels strike the right balance between accuracy and usability, to become the new standard ship performance model for operational optimizations like routing, maintenance planning (fouling detection) and speed optimization.

3. Why sensor data is a must for accurate ship performance modelling

In the tables below, we compare how well Noon Report (NR) data approximates High-Frequency Data (HFD) measured by sensors. The goal is to understand the measuring error by NR data for parameters like Speed Through Water (STW) and Main Engine Power. In the first scenario we consider HFD averaged over 24 hours to be the ground truth (1 averaged point per day). In the second scenario, we consider the HFD averaged every 5 minutes to be the ground truth (288 points per day). The error is expressed as the Mean Absolute Percentage Error (MAPE).

The first column shows the error metrics in the case NR data is compared to HFD data averaged over 24h (the timespan covered by the NR). The second column takes the un-averaged HFD data and compares it to the corresponding NR by assuming the NR data is valid for the HFD datapoint that falls within its covered timespan. Note that this increases the error as we are effectively upsampling or interpolating the NR data to the HFD frequency. The second column shows the added value of HFD data while the first mainly shows the effect of manual corrections or wrong entries in NRs.

MAPE - STW	**NR compared to 24h averaged HFD**	**NR compared to 5 min averaged HFD**
Ship 1	0.9%	4.1%
Ship 2	1.1%	4.1%
Ship 3	1.7%	3.1%
Average	**1.2%**	**3.8%**

MAPE - Power	NR compared to 24h averaged HFD	NR compared to 5 min averaged HFD
Ship 1	2.6%	7.6%
Ship 2	1.9%	8.1%
Ship 3	13.2%	14.0%
Average	**5.9%**	**9.9%**

In the second column it can be observed that NR data has an average error of about 3.8% for STW and an error of 9.9% for power. It can be seen that the power shows large differences between NR and HFD data. We can conclude that using NR data instead of HFD measured by sensors adds a significant error to STW and Power. This inaccurate measurement severely limits the potential to create and validate ship performance models. In the absence of HFD measured by sensors, NR data is too inaccurate to be used as a ground truth to build and validate models with a power prediction error of less than 10% (MAPE).

4. Methodology

4.1 Measuring Accuracy for Ship Performance Modelling

We advocate for the next 3 metrics to become industry standard, given they are comparable over multiple ships and can be linked directly to certain operational optimization use-cases.

R^2 = Determination Coefficient

How much of the variance in power is explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Use:

-Estimating the goodness of fit.

MAPE = Mean Absolute Percentage Error

The relative error per single prediction expressed as a percentage. Given we consider high-frequency data on power at 5 minute intervals as the ground truth, this is the error made per 5 minute interval.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Use:

-Operational Optimizations like weather routing and speed optimization that make a trade-off between a wide variety of operating conditions, also considering shorter time periods

-Short term performance monitoring to quickly identify severe underperformance issues

MAMPE = Mean Absolute Monthly Percentage Error

The error in estimating the performance averaged over a month of sailing.

MAMPE is similar to MAPE, but has the crucial difference that instead of averaging the absolute relative error for all data points, it calculates the average relative error per month of sailing before taking the mean of the absolute values. This allows over- and underestimations due to sensor measuring volatility to even out.

$$\mathrm{MAMPE} = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{1}{n}\sum_{j=1}^{n}\frac{y_{ij}-\hat{y}_{ij}}{y_{ij}}\right|$$

Note that the first summation indexed by i loops over months (m) while the second summation indexed by j loops over n data points within a month.

Use:

-Performance monitoring and fouling detection over the longer term

4.2 Modelling Scope

ML models are trained and validated on different separate datasets.This ensures the accuracy metrics represent the performance of the models for unseen data, as this is the way the models would be used in practice. When calculating the accuracy, the sensor data is considered as the true value, even if the model is only trained on noon report data. It is expected that sensor data is a more reliable ground truth than noon report data, if the sensors are well calibrated and checked for outliers.

To set a correct baseline for a “normal mode of operation” for the vessel the training data is selected to be in time intervals closely following a dry-docking. This allows for an accurate estimation of the power overconsumption and speed loss due to fouling, hull degradation, and other time-dependent factors that impact the performance of the vessel.

It is crucial to recognize that ship performance modelling consists of multiple conversions or modelling steps. Some steps are straightforward and can be well approximated by empirical formulas (speed over ground (SOG) to speed through water (STW), power to fuel consumption). For the STW-RPM-Power relationship however, significant accuracy increases can be reached by using machine learning compared to a combination of sea trial curves and correction formulas. This is due to the high-dimensionality and high complexity of the relation, an ideal challenge for Machine Learning algorithms when sufficient high-quality data is available.

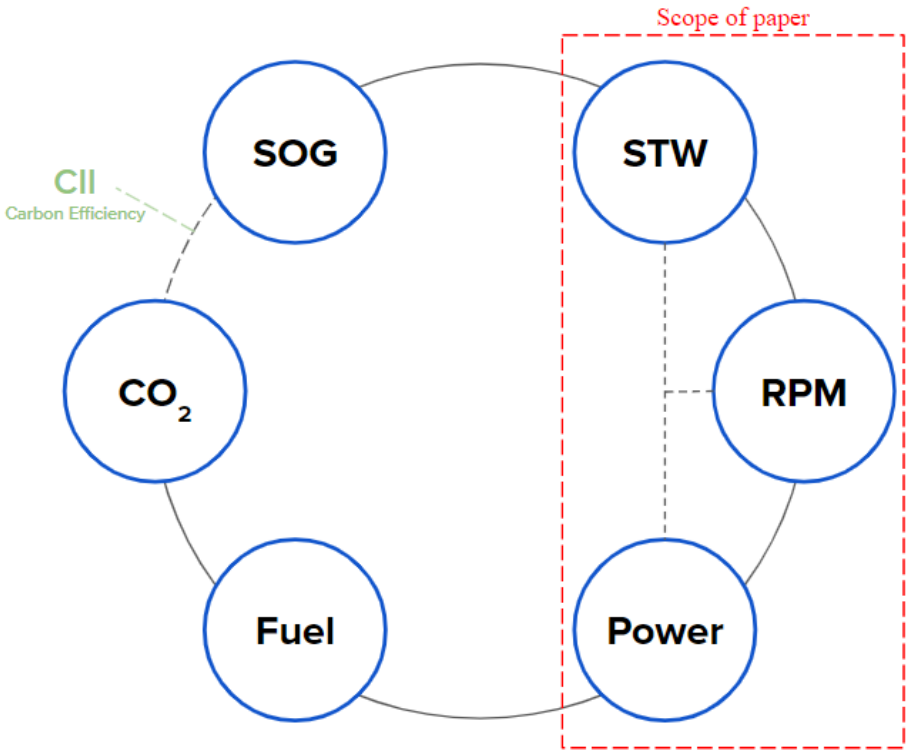

Figure 2: Sub-relations in a Ship Performance Model, red zone indicating the scope of this paper. Note that the accuracy metrics are not necessarily equal in both directions[1].

This paper focuses on the conversion from STW to power. This relation is the most difficult to model accurately and its accuracy dictates how well hull performance can be analysed.

4.3 Modelling Approaches

In order to quantify the improvement of the ML models, its predictions are compared to other approaches, ranging from a simple baseline model to more advanced semi-empirical formulas.

A baseline model widely used in the industry are sea trial curves or model tests. This is a simple function translating STW to power which does not take any other operational factors or weather conditions into account. This method is used to set a baseline to compare the other approaches with. Although this method can be regarded as an oversimplification, it is sometimes used in the industry.

A common approach to improve the baseline sea trial curves or model tests consists of correcting the power prediction for weather factors such as wind and waves using formulas. Here we use the ISO15016 standard [5] for wind correction and Kreitner's method [6] for wave correction.

[1]For example suppose that a range of 10% in STW corresponds with a variation of only 5% in RPM. Predicting STW->RPM in this range will lead to errors of at most ~5% as there is only 5% variation in the target, while the inverse relation, RPM->STW can have errors of up to 10%.

Next, a Machine Learning model trained on noon report data, enriched with third-party weather data, is investigated. The ML model is trained to predict the main engine power using weather information and vessel conditions such as stw, draft,... Note that noon report data is only entered once every 24 hours and is subject to human errors. As such it can be expected that these models are still far from the best solution possible.

Finally a new ML approach denoted “Ship Kernels” is detailed. The Ship Kernels are trained using sensor data combined with weather data. The regression task is basically the same as the NR models but as much more data is available, more ML solutions, such as neural networks (NN) become feasible. Several known physical relations from naval engineering are enforced to create physically consistent models. This can be achieved using Physics-Informed Machine Learning [7]. It is a highly non-trivial task and is one of the major strengths of our models compared to other data-driven solutions.

5. Results

Figure 3 shows the “learning curve” for the 4 modelling approaches outlined above. A learning curve shows how the accuracy of the model changes as more data becomes available. The sea trial curve and the “sea trial + correction” approach are not fitted to any data, leading to horizontal lines on the learning curve figure. For data-driven techniques like machine learning, the amount of training data has a large influence on the accuracy. Both data-driven methods improve in accuracy until they stabilise after approximately 5 months of training data.

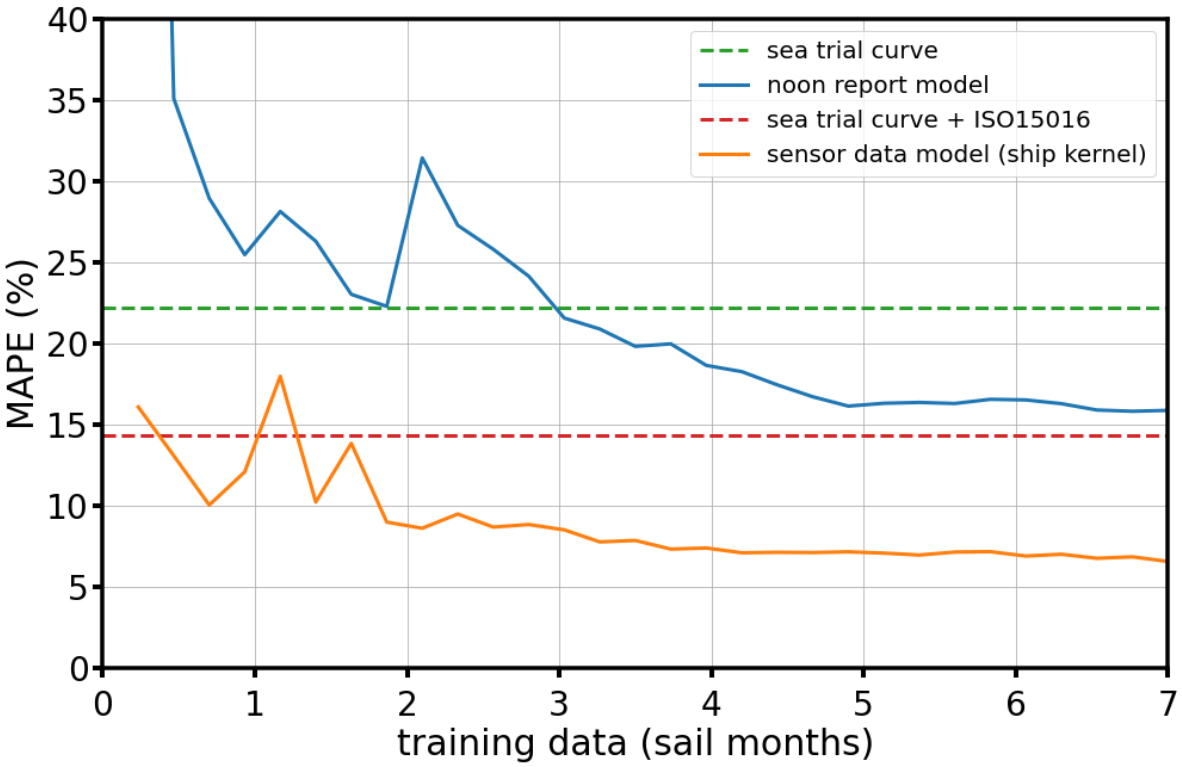

Figure 3: The learning curves using MAPE for the different approaches outlined in this paper.

Below the accuracy metrics for all 4 modelling approaches are listed. Ship kernels have the highest goodness of fit, followed by 'sea trial + corrections'. The NR model trained on NR data and validated on sensor data has a very bad goodness of fit, possible due to misreporting and inaccuracies in the NR data. Sea trial curves are the least accurate approach, as was to be expected from an approach that can't account for changing conditions.

STW to Power	**Sea trial**	**NR model**	**Sea trial + correction for wind & waves**	**Sensor data model (Ship Kernel)**
R^2	0.38	0.26	0.73	0.86
MAPE	22.2%	15.9%	14.3%	6.7%
MAMPE	21.9%	13.6%	13.7%	2.6%

It can be observed how the standard 'Sea trial + correction'-approach has a comparable accuracy to ship kernels when only NR data is available (14% and 16% MAPE respectively). However, in a scenario where sensor data is available, the ship kernels have the best accuracy metrics. With a MAPE of 6.7%, the ship kernel is more than double as accurate as the 'sea trial + corrections'. Investigating the MAMPE shows that the data-driven approach of ship kernels drastically outperform other approaches, making it a much more accurate option to analyse long-term ship performance related to hull & propeller fouling.

6. How to increase the operational usability of Machine Learning

A disadvantage of Machine Learning for ship performance modelling is that it requires 3-6 months of operational data before an accurate model can be made [8]. A second barrier is the requirement of sensor data, while the majority of ships only have NR data today.

Toqua has solved this disadvantage by creating models that draw from prior knowledge learned by ship kernels for vessels of similar design with sensor data[2]. This approach is denoted as the "augmented approach" and its learning curve is displayed in figure 4. It can be observed how a MAMPE of around 7% is possible for a ship that only has NR data. This makes this 'augmented approach' the most accurate solution for vessels that only have NR data, outperforming the 'sea trial + corrections'-approach that has a MAMPE of 16%. Nevertheless, to reach the highest accuracy sensor data is still required. We argue that the additional accuracy definitely merits the investment cost in sensor data, since more accurate performance understanding leads to better decision making and even the smallest relative fuel savings outweigh the absolute investment cost of sensor data.

[2]A detailed explanation of this method is deemed to be outside the scope of this paper

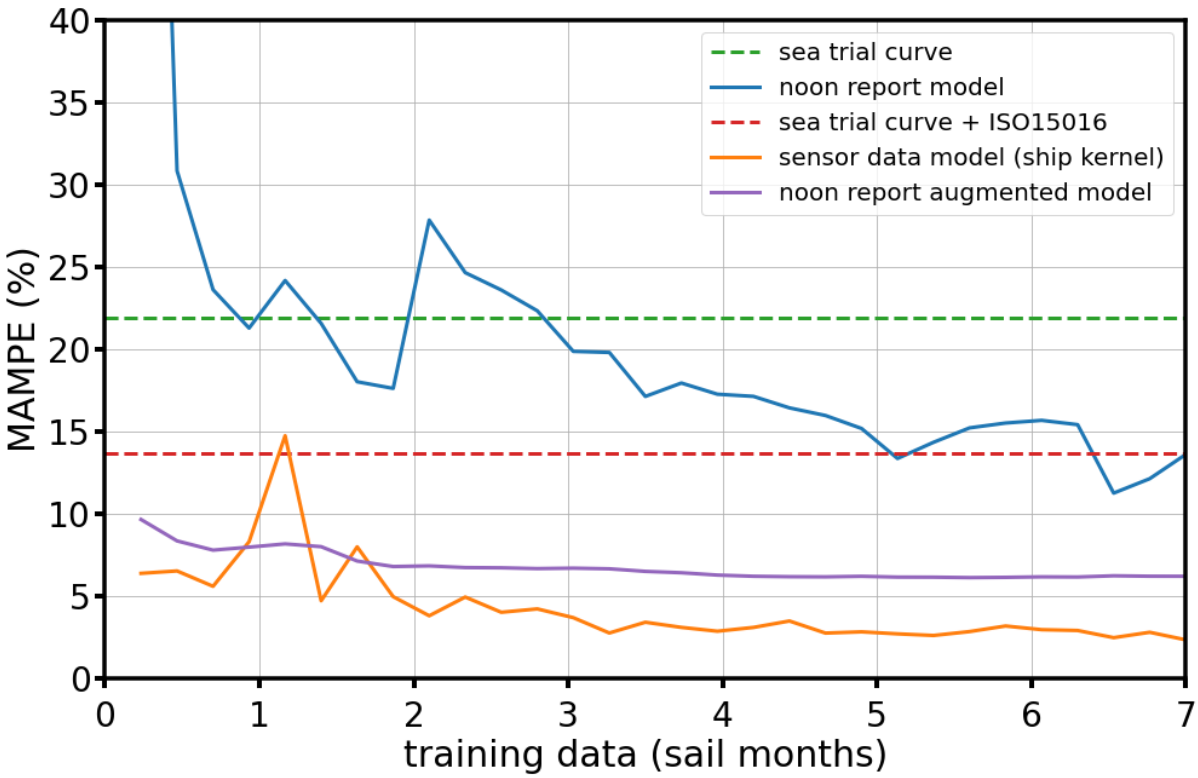

Figure 4: The learning curves using MAMPE for the previous approaches and the 'augmented approach'

7. The real challenge of using ML in shipping

According to Gartner [9], 85% of ML projects fail. We expect that number to hold true in the shipping industry as well. The real challenge does not lie in creating the most accurate and sophisticated models but in getting those models operational at scale in a cost-friendly and reliable way. Most ML projects die after a Proof-Of-Concept stage. A machine learning model can quickly show promising results, but putting that model into production and serving it to the world adds many new challenges and development costs causing projects to lose momentum, go over budget and eventually be cancelled.

In the evolution of ship performance modelling, the first challenge is gathering high-quality, high-frequency data. Today more and more companies with sensor data have reached a stage where their data is ML-ready. The second challenge is to create accurate and robust models from this abundance of data. As this paper illustrates, physics-informed Machine Learning models like Ship Kernels can deliver on that promise. The final and largest challenge before the benefits of digitalization in shipping can really be achieved is getting these models operational at an industrial scale. The costs, time and people required to get ML in production (MLOps) are a multiple of the resources required to create a Proof-Of-Concept model [10] [11] [12]. We view this as the largest challenge the industry will be facing in the next few years in achieving the efficiency-gains promised by operational optimizations powered by digitization and better ship performance modelling.

8. Summary

A qualitative analysis is presented of the ship performance modelling techniques available to the industry today. Current techniques are graded on their suitability for operational optimizations, requiring high amounts of predictions for a wide range of combinations of speed, draft and weather conditions. A case is made in favour of techniques drawing from a combination of physics-driven and data-driven insights.

Secondly, the measuring inaccuracy of STW and power for Noon Report data is quantified by comparing it with High-Frequency Sensor data. A significant measuring error is found due to Noon Report data (MAPE = 3%-14%). This prompts the authors to conclude that sensor data is an undeniable requirement in order to create and validate highly accurate data-driven ship performance models.

Next, a quantitative analysis is made comparing the accuracy of different modelling approaches for the conversion from STW to ME-Power. It is found that ship kernels (ML, developed by Toqua) trained on sensor data have a much lower error (MAPE=6.7%) compared to other approaches (Sea trial curves: MAPE = 22%, 'sea trial + wind & wave correction': MAPE = 14%, ML on noon report data: MAPE = 16%).

Finally, a fair warning is given that the real challenge in capturing the value of ML & sensor data, does not lie in creating accurate models, but in getting these models operational at an industrial scale, in a reliable and cost-effective manner (MLOps).

Bibliography

[1] "Forty percent of 'AI startups' in Europe don't really use AI, says report - The Verge." https://www.theverge.com/2019/3/5/18251326/ai-startups-europe-fake-40-percent-mmc-report

[2] "The First Rule of Machine Learning: Start without Machine Learning." https://eugeneyan.com/writing/first-rule-of-ml/ (accessed Apr. 17, 2022).

[3] D. W. Taylor, *The Speed and Power of Ships: A Manual of Marine Propulsion*. J. Wiley & Sons, 1910.

[4] J. Holtrop and G. G. J. Mennen, "An approximate power prediction method," *International Shipbuilding Progress*, vol. 29, no. 335, pp. 166–170, 1982.

[5] "ISO - ISO 15016:2015 - Ships and marine technology — Guidelines for the assessment of speed and power performance by analysis of speed trial data." https://www.iso.org/standard/61902.html (accessed Apr. 17, 2022).

[6] K. J., "Heave, Pitch and Resistance of Ships in a Seaway," *Trans. INA*, vol. 81, no. 440, p. 203, 1939.

[7] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nat. Rev. Phys.*, May 2021, doi: 10.1038/s42254-021-00314-5.

[8] B. Volker, "6th Hull Performance & Insight Conference," in *6th Hull Performance & Insight*

Conference, 2021.

[9] "Gartner Says Nearly Half of CIOs Are Planning to Deploy Artificial Intelligence." https://www.gartner.com/en/newsroom/press-releases/2018-02-13-gartner-says-nearly-half-of-cios-are-planning-to-deploy-artificial-intelligence (accessed Apr. 17, 2022).

[10] E. Clemmedsson, "Identifying pitfalls in machine learning implementation projects: a case study of four technology-intensive organizations," 2018.

[11] "Estimating the Time, Cost, & Deliverables of an ML App." https://appinventiv.com/blog/machine-learning-app-project-estimate/ (accessed Apr. 17, 2022).

[12] I. Lee and Y. J. Shin, "Machine learning for enterprises: Applications, algorithm selection, and challenges," *Bus. Horiz.*, Nov. 2019, doi: 10.1016/j.bushor.2019.10.005.