On the Evaluation of Uncertainty of AI models for Ship Powering and its effect on Power Estimates for non-ideal conditions.

Efthymia Ofelia Tsompopoulou, DeepSea Technologies, Athens/Greece, E.Tsompopoulou@deepsea.ai
Andreas Athanassopoulos, DeepSea Technologies, Athens/Greece, A.Athanassopoulos@deepsea.ai
Elli Sivena, DeepSea Technologies, Athens/Greece, E.Sivena@deepsea.ai
Kyriakos Polymenakos, DeepSea Technologies, Athens/Greece, K.Polymenakos@deepsea.ai
Vasileios Tsarsitalidis, DeepSea Technologies, Athens/Greece, V.Tsarsitalidis@deepsea.ai
Antonis Nikitakis, DeepSea Technologies, Athens/Greece, A.Nikitakis@deepsea.ai
Konstantinos Kyriakopoulos, DeepSea Technologies, Athens/Greece, K.Kyriakopoulos@deepsea.ai

Abstract

Accurate modelling of vessel behaviour has never been more important in the shipping industry. While data-driven methods and deep learning approaches are rapidly gaining popularity, the lack of an established and universal process for assessing the accuracy of a vessel's model is a significant obstacle to widespread adoption throughout the sector. In this work, we present an evaluation methodology, based on a dataset-splitting scheme, that aims to reveal a given model's robustness or deficiency in the face of the distributional shifts that inherently characterise many maritime datasets. As part of this process we examine the results through the lens of predicted uncertainty, yielding useful information about a model's fitness in dealing with uncertain and noisy regions in the modelled dataset. We introduce a realistic synthetic dataset allowing for the systematic study of a model's performance under a distributional shift. In the proposed dataset: a) we model all aspects of drag as described by the literature, including the rate of hull fouling through time; b) our input features are real samples from on-board sensors collecting data at high frequency; and c) we inject different patterns of noise to measure the model's predictive uncertainty performance. The outcome is a repeatable and robust methodology that allows for the assessment of vessel performance models within the context of their used environment - having a direct impact on a models' deployability and, ultimately, on their ability to deliver meaningful insights.

1. Introduction

Environmental regulations, fuel prices, and societal factors all push stakeholders to work even harder to reduce the shipping industry's carbon footprint. At the moment, simply doing their best and/or purchasing equipment labelled "eco" is insufficient. Vessel performance must be measured and evaluated on a continuous basis to establish changes over time and benchmark against the global fleet. With the introduction of EEXI and CII, as well as emissions trading schemes, every vessel will need to be set on a path of continuous improvement, as well as objective measurement of performance. To this end, each stakeholder will have to employ a variety of tactics, ranging from technical retrofits to alternative fuels, condition-based maintenance, and operational measures. One of the most accessible means of boosting performance is to develop a new layer of understanding about what really influences vessel performance.Modelling the real-world performance of the hull and propeller will allow for safer decisions on timely cleaning. Furthermore, such modelling can aid in the evaluation of other investments (e.g., antifouling paint, ESD retrofits, etc.) and provide a clearer image of each vessel's operational capabilities, such as charter party description or bunker budgeting. Using the same data, legislators and authorities can estimate the environmental impact of best practices and thus plan future regulations. With detailed vessel performance models in hand, operators can optimise a ship's route and speed profile to reduce fuel consumption. Such a strategy has been widely recognized as the most immediate way of making an impact on emissions, with potential for remarkable results yet requiring low up-front investment. However, the extent to which this approach translates into realised emissions and fuel savings is highly dependent upon the accuracy and granularity of the models that underpin it.

All of the above strategies necessitate an objective method for measuring their impact on reducing fuel consumption and tracking progress in decarbonization efforts. With the advent of ISO 19030 *(2016)*, a standard that levels the playing field was introduced, and a reference frame to follow was set for everyone who cares about their vessels' performance. A good framework was set, consisting of minimum required data to be collected, and criteria for outliers and basic filtering (even though it can be very restrictive). In the core of ISO 19030 the major point of strength, but also vulnerability, is the Reference Model, which (for now) can only be constructed using towing tank tests, sea trials and "detailed" CFD calculations. As discussed in previous papers *(Tsarsitalidis & Rossopoulos 2018)*, special attention is needed to the quality of the model, which makes its production more expensive for the owner, while also maintaining a series of potential issues. Missing Hull, Propeller and Appendages information, push the model builder to make assumptions with serious impact and possibly several trial and error iterations until a safe model is reached, if at all.

Hull interaction factors can be difficult to estimate, especially in the case of retrofits, while even if experimental data are available they are usually collected for a narrow set of conditions (speeds, drafts) and very rarely simulated for rough (i.e non-calm) sea. Additionally, the effects of swell are completely disregarded. Even when everything is executed perfectly, the ISO provides a reliable and objective evaluation of performance change only in good weather and steady conditions. It is safe to assume that an improvement in hull roughness and drag measured in good weather will remain an improvement in rough seas or unsteady conditions, but it cannot be taken for granted that any hydrodynamic retrofit will maintain its good characteristics in non-ideal conditions. Furthermore, the current version of the ISO does not permit such testing to confirm or disprove any technology.

In response to such problems, data driven methods are on the rise, where deep learning is showing serious promise *(Górski et. al. 2021, Levantis et. al. 2020, Gonzalez et. al. 2019 Park et. al. 2018)* , but even these are not void of limitations and potential problems. A fundamental issue with such models is the lack of formal performance guarantees that would specify sufficient conditions before training for the models to reach a certain level of performance. Consequently the accurate model evaluation after training but before deployment is of paramount importance. In practice, these complex models are usually evaluated on some part of the available dataset that has been held-out during training, based on one or more simple error metrics, such as the root mean square error and the mean absolute percentage error.
The underlying assumption of this standard evaluation practice is that the training, validation and deployment datasets are independent and identically distributed (i.i.d.) and thus testing on the available dataset is indicative of performance at the deployment phase. Unfortunately, this assumption does not hold in real world applications, where the data distribution of the training and the validation set shift from the distribution the model faces when finally deployed *(Malinin et al. 2021)*. Recently, significant research effort has been invested into the accurate evaluation of ML models, especially in high risk or sensitive contexts *(Amodei et al. 2016)*.
Furthermore, even assuming that we have a unique and accurate numerical estimate quantifying the average predictive accuracy of a model, there is still essential information missing. A model that is overall fairly accurate can be wildly inaccurate in a relevant subset of the dataset, and this mismatch between the on-average and the worst-case performance can lead to catastrophic down the line decision making (e.g weather routing). The notion of predictive uncertainty, where the model provides not only a prediction, but also an estimate of how confident it is for that particular prediction, allows for more nuanced risk analysis and more effective planning. Although there are a plethora of methods for estimating predictive uncertainty that can be compatible with deep neural networks, a solid evaluation procedure is needed to assure that the researched models are producing well calibrated uncertainty estimates.

Supporting this effort, we introduce a synthetic dataset [1]along with a well-thought split that could help in the direction of systematic evaluation of deep learning models. We show that predictive uncertainty can reveal important information regarding the distributional shift between training and testing but also regarding the dataset splits and the noise levels of the target signal. Although working with a synthetic dataset we are not limited by it, since the same process can be transferred in real datasets. We hope that this work will positively impact the shipping industry into trusting and adopting deep learning methods with predictive uncertainty in vessel performance modelling.

2. Background

In recent years, Machine Learning and Artificial Intelligence (AI) methods have achieved state-of-the-art (SOTA) performance utilising large amounts of data, becoming the default option to solve fundamental problems in various domains such as computer vision *(Krizhevsky et al. 2012)*, speech recognition *(Hinton et al., 2012)*, natural language processing *(Mikolov et al. 2013)* and bioinformatics *(Alipanahi et al. 2015; Zhou & Troyanskaya 2015; Ramsundar et al. 2015)*. These advances have brought an increasing number of practical production level autonomous decision systems with high-risk outcomes for their users, such as in finance, medicine, autonomous vehicles and in shipping *Coraddu et al. (2019)*.

The key argument in favour of ML methods is that while traditional methods rely relatively more heavily on expert prior knowledge and expensive computation at inference time, i.e. whenever predictions are needed, ML methods utilise more efficiently larger datasets and computational resources at training time, which can be done offline, with relatively smaller prediction costs. Given the clear trend towards larger and larger datasets, and the development of increasingly sophisticated computational resources dedicated for ML applications, a strong case can be made for the future of ML methods in shipping. While a plethora of ML methods have been presented, each with its unique pros and cons, at the current state of ML research, deep neural networks tend to be the default choice for a great number of tasks, especially those with abundant, unstructured datasets.

Machine learning in shipping is relatively new. A good review of early attempts can be found in *Coraddu et al. (2019)*. A machine learning approach to outlier detection and filtering was shown by *Gonzalez & Arango (2019)*, where it was proven to be extremely difficult to distinguish between anomalies and ship operation induced bias (despite the use of high quality sensors and very low overall noise), while the need of ship specific parameters in filtering was recognized. *Gorski et. al. (2021)* displayed the potential of unsupervised learning algorithms, by means of clustering and identifying the most frequent modes of operation of a vessel and building the baseline model for these conditions, with the obvious limitation of being restricted to the specific modes of operation. Thus, the problem of reliable generalisation remains and uncertainty modelling is possibly our best way of measuring (and then improving) the quality of our models.

2.1 Uncertainty and Modelling

In machine learning there are two distinct types of uncertainty that can be modelled: aleatoric and epistemic uncertainty *(Kiureghian & Ditlevsen 2009)*, while the term total uncertainty refers to their sum. Aleatoric uncertainty captures the inherent stochasticity of the problem, with the rolling of a dice being the prototypical example. This type of uncertainty is not reduced with additional data, as the extra data do not affect the stochastic nature of the problem. Epistemic uncertainty on the other hand captures the uncertainty introduced by incomplete information about the data generating process. As we obtain more data, epistemic uncertainty can be reduced.

[1] Dataset will be distributed from SHIFT 2.0 challenge competition (http://deepsea.ai/datasets)

There has been significant effort in recent years to combine the power of deep neural networks with the probabilistic techniques that allow for uncertainty estimation and decomposition. Various terms have been used in the research community: Bayesian deep learning (BDL) *(Wang and Dit-Yan 2020)* usually refers to a general framework that combines probabilistic thinking with deep learning, that sometimes includes self-supervision and active learning, while Bayesian Neural Networks (BNNs) *(Goan and Fookes 2020)* usually refer to deep neural networks augmented with appropriate techniques to support uncertainty quantification. The most popular methods are based either on variational inference, dropout, or ensembles of networks *(Lakshminarayanan et al. 2017)*.

2.2 Generalisation and Distributional Shift

Training a ML model usually involves minimising an error metric on the available dataset (empirical risk minimization, ERM), while the actual goal is to minimise the expectation of the error over the unknown data generating distribution, also referred to as the risk *(Hardt & Recht 2021)*. The generalisation gap is defined as the difference between the empirical error and the risk, and intuitively is the difference in performance of the model on the data it has been trained on, compared with unseen data from the same distribution. Distributional shift takes the notion of generalisation a step further, taking into account the model's performance when evaluated on unseen data coming from a different data generating distribution, in comparison to its performance on unseen data from the same distribution as the training data. It's worth noting that robustness to arbitrary distributional shifts is impossible: the two distributions have to be in some sense similar for the model to perform well (for a detailed breakdown see *Moreno-Torres et al. 2012*). Both a large generalisation gap and a large performance degradation due to distributional shift are often interpreted as proof that the model is overfitting. This is partially accurate, since not measurable (within available dataset) overfitting is a necessary but not sufficient condition for good generalisation and robustness. Various inherent biases in the sampling process due to extended missing rate or sailing in specific conditions can significantly bias the model without being detectable within the given dataset by using the common model cross validation procedures. The analysis of such selection biases and to what extent they could affect the model's generalisation ability could be the focus of future work.

Under this scope, we propose an evaluation methodology built around a carefully designed dataset partitioning scheme aiming at exposing a model's robustness to large distributional shifts. For the purposes of this study the dataset is synthetic but the same methodology could be transferred to real ones. As part of the proposed methodology, we analyse the results through the lens of predictive uncertainty as it can reveal useful information about the model fitness in handling uncertain and noisy regions in the modelled dataset. In order to make the dataset as realistic as possible: a) we model all aspects of drag as dictated by the naval engineering literature including hull fouling effects over time b) our input features are real samples from high frequency on-board sensors (i.e real seeds), augmented with real weather data and c) we inject different patterns of noise to measure the performance of model's predictive uncertainty.

3 Methodology

3.1 Synthetic dataset

In this work we introduce a synthetic dataset based on real vessel's seeds and realistic noise to train and evaluate machine learning models. There are multiple distinct advantages that come with this choice, along with certain drawbacks. First and foremost, a synthetic dataset provides access to the ground truth values of all data points. These values can be corrupted by noise if needed, e.g. to test the model's robustness to noisy training data, while the targets without noise can still be used for

evaluation. Additionally, the availability of the ground truth labels gives us complete control over the introduced noise both in terms of its magnitude (as we can control the signal to noise ratio) and its properties (uniform white Gaussian noise, heteroscedastic noise etc.). The synthetic dataset also allows us to create as much data as needed (will be explored in a future work), distributed according to the demands of each particular experiment, while a real dataset would only allow choosing a subset of its data points to simulate such effects. Finally synthetic datasets can eliminate data confidentiality concerns, and as a result promote the sharing of methods and results. On the other hand, synthetic datasets may deviate from real world data, by failing to simulate realistic scenarios or inaccurately portraying some of their properties, casting doubts on whether the conclusions will hold in practice.
Since the focus of this work is to systematically highlight possible complications with the deployment of neural network models through a well-considered dataset partitioning, the need for a well-controlled synthetic dataset was necessary. This is not limiting, since one could apply the suggested dataset partitioning along with the proposed predictive uncertainty measures to real datasets as well.

3.2. Dataset construction

3.2.1 Synthetic model

The Synthetic model is a generative function ($f_{synthetic}$) taking as input a time-series of features (i.e. signals), as recorded from a real vessel, and calculates the power consumed by the vessel's hull. This function finds the propeller cooperation point after calculating all the components of resistance (bare hull, appendages, wind, waves, fouling drag) for given speed, draft and trim. More specifically, for the generation of synthetic data, a non-linear solver script was created to find the operating point of a given propeller and hull resistance for each desired condition, as described by *Bose (2008)*. The propeller curves (K_T, K_Q) can either be user defined or use the B-Series *(Van Lammeren et al. 1969)*. For the resistance part, the calculation of each component can be described as follows: having the full hydrostatics table of the vessel for the whole range of drafts and trims, along with a series of geometric characteristics (bulb shape and size, transom, appendages etc), calm water resistance is calculated by employing the Holtrop method for slender ships (i.e. containers, RoRo, gas carriers) and Modified Holtrop *(Nikolopoulos & Boulougouris 2018)* is used for bulkier ships like large Tankers and Bulk carriers. Following the ISO 15016 *(2015),* the weather added resistance is found by calculating the Wind effect by using the the regressions of *Fujiwara et al. (2006)*, while the wave effects are modelled according to STAwave1 and STAwave2 as also introduced by *Tsujimoto et al (2008)*. Hull Interaction factors are calculated depending on ship type, using empirical formulas, a summary of which can be found in *Carlton (2018)*. Scale effect corrections, cavitation criteria and corrections were also taken from *Carlton (2018)* and *Bertram (2012)* . The effect of wake affecting energy saving devices can be modelled by adjusting the interaction factors. Fine-tuning of the method to fit a specific vessel (when there is not enough hydrostatic data, or discrepancies are observed), can be done by using sea trial data and/or detailed factors when available from a towing tank report, or actual measurements of well known conditions. Last but not least, the effect of fouling is modelled as the result of its manifestations (drag, propeller and interaction). The change in drag coefficient is modelled after *Townsin (1981)*, the effect of fouling on the propeller performance is modelled as in *Seo et al. (2016)* (increase in torque coefficient), as also described in *Carlton (2018)* and the change of interaction factors are modelled after *Farkas et al. (2020)*. All the aforementioned models produce the effect of fouling on each component over time, which is measured from each drydock / cleaning event.

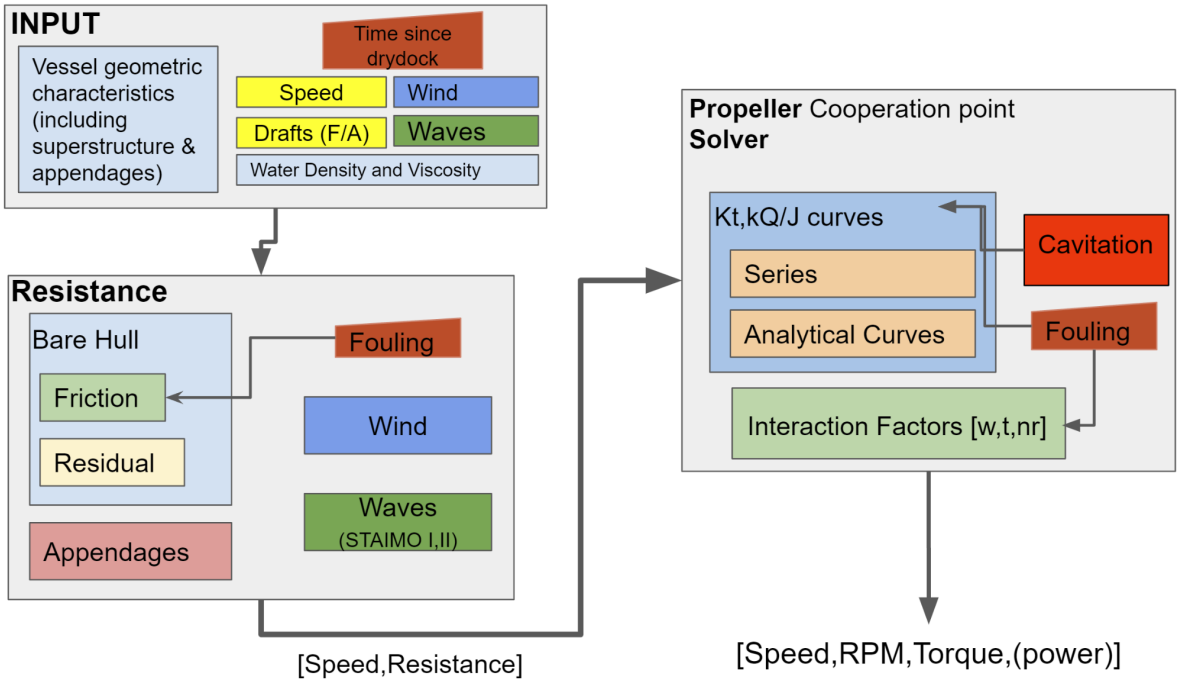


Fig.1 : Structure of the synthetic model. The model consists of two main parts, resistance and propulsion solver. The first part (resistance) calculates all the components of resistance for given speed, drafts and weather conditions based on the ship characteristics and time. Then, the Propulsion solver finds the cooperation point for the calculated resistance and speed, with given propeller characteristics and interaction factors.

The resulting program is depicted in Fig.1 and given the characteristics and data of an actual ship, it can estimate the power, rpm and torque, for any given combination of Speed, Draft, Trim, weather conditions, and time since the last drydock, within the limitations of the methods used. The primary vessel particulars are given in Table I for reference.

Table I: Ship Characteristics

Metric name	Value
type	Bulk Carrier
Length Overall	292 m
Length BP	282 m
Breadth (Mld)	45 m
Depth (Mld)	24.8 m
Design Draft	16.5 m
Scantling Draft	18.3 m
Deadweight (at Tdesign)	176364 Tons
main engine MCR	16860 kW
Design speed	14 kn
Operating Speed	13.5 kn

3.2.2 Dataset features

For the current investigation, the dataset is created by combining the real samples with synthetic power labels generated by our synthetic model as described in subsection 3.2.1. The real vessel's records have been sampled on a per minute basis covering a time period of more than four years. The available features as presented in Table II, are recorded by on-board sensors, the global positioning system (GPS) and are augmented with weather data from a global weather provider. The data is preprocessed to remove stationary states, for example when a vessel is at port.

Table II: Available features of synthetic set.

Feature name	Units	Description
Synthetic power	kW	Synthetic propeller shaft power (Target)
Draft aft	m	Vessel's draft at stern from noon reports
Draft fore	m	Vessel's draft at bow from noon reports
Stw	kn	Speed through water
Acceleration	Kn / 3 min	Acceleration over ground
Apparent wind speed	kn	Apparent wind speed
Apparent wind vcomp	kn	Apparent wind component along vessel's direction of motion
Apparent wind ucomp	kn	Apparent wind component perpendicular to vessel's direction of motion
Recurrent vcomp	kn	Relative current component along vessel's direction of motion
Recurrent ucomp	kn	Relative current component perpendicular to vessel's direction of motion
Combined waves height	m	Combined wind (sea) and swell wave height
TimeSinceDryDock	min	Time feature quantifying the time period from the last Dry Docking cleaning event

3.2.3 Dataset Partitioning

The focus of this work is to help with the development of robust models with high quality uncertainty estimates on distributional shifts. Under this scope and motivated by the canonical partitioning of the weather dataset presented in *Malinin et al. (2021)*, we split the synthetic set in two dimensions: time[2] and true wind speed as illustrated in Fig.2, using the wind speed intervals of Table III. The time dimension aims to capture the non-stationary effects of fouling while the wind speed dimension aims to capture weather effects (by being a proxy since wind is correlated with wind-waves) and to better expose the model's performance in bad or uncertain weather. Partitioning the dataset in more dimensions would have added complexity without practical benefits since the most important factors of uncertainty (weather and fouling) are already represented.

The proposed partitioning has the primary goal of assisting in evaluating the true performance of a model given a real dataset as nearly as possible. To systematically demonstrate its efficacy for this study, we had to employ a synthetic but nonetheless realistic dataset, allowing us to preserve full

[2] No cleaning events take place during the time period covered by the synthetic set.

control over the dataset properties while also having access to ground truth labels.

Three main subsets are created from the proposed partitioning: the train set, used for model training, and the development and evaluation sets, used for the evaluation of the model performance.
In more detail:

- **Train set**: It covers the time range of 39.4 months starting after a dry docking cleaning event and includes data with true wind speed up to 19 kn.
- **Development set**: It consists of an in-domain partition dev_in and an out-of-domain partition dev_out, with equal representatives achieved by downsampling dev_out to match the number of records of dev_in. Dev_in is sampled from the same partitions as the train set while dev_out includes more recent records (time period of 6.6 months) that correspond to wind speeds in the range [19, 26) kn.
- **Evaluation set**: Same as for development set, evaluation set has an in-domain eval_in and an out-of-domain partition eval_out having equal populations (eval_in is downsampled in this case). Eval_in is sampled from the same subsets as the train set. Eval_out is the most shifted partition from the in-domain distribution, including the most recent records covering a time period of 18 months and the most severe wind conditions encountered in the dataset, corresponding to wind speed range [19, 40) kn.

Table IV shows the number of records of the proposed partitions (rows) along with the respective populations in each 2D segmentation of the synthetic set (columns with prefix group).

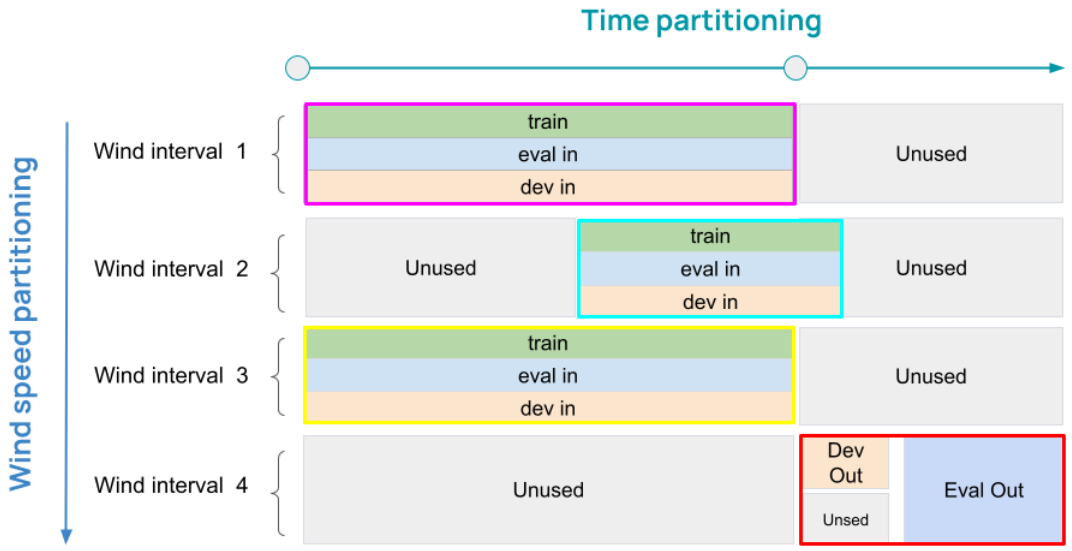


Fig.2 : Canonical partitioning of synthetic set.

able III: Wind intervals considered for data partitioning. Beaufort ranges are defined approximately.

Wind interval	Range (kn)	Range in Beaufort
1	[0, 9)	Up to ~3
2	[9, 14)	3-4
3	[14, 19)	4-5
4	≥19	≥ 5

Table IV : Number of records in the canonical partitioning of synthetic set. The colored borders of the group columns indicate the dataset segmentations from which the partitions are sampled following the same color notation as in Fig.2.

Data	pct (%)	total	Group 1	Group 2	Group 3	Group 4
train	80.3	523190	231626	118698	172866	0
dev_in	-	18108	8017	4108	5983	0
dev_out	-	18108	0	0	0	18108
dev	5.6	36216	8017	4108	5983	18108
eval_in	-	46021	20355	10448	15218	0
eval_out	-	46021	0	0	0	46021
eval	14.1	92042	20355	10448	15218	46021

3.2.4 Target noise

One of the primary goals of this work is to investigate the quality of uncertainty estimation both within and outside of domain areas. Working with a synthetic dataset enables well-controlled noise pattern injection, which should be captured by the model's heteroscedastic predictive uncertainty. We apply two types of Gaussian noise with non-constant variance (heteroscedasticity) to the synthetic target y_i to make the synthetic set realistic for this task:

- heteroscedastic Gaussian noise correlated with power, $\varepsilon_{power,\,i} = N(0,\, a{\cdot}y_i)$. This type of noise simulates the scenario of linear deterioration of the torque meter accuracy as power increases,
- heteroscedastic Gaussian noise correlated with true wind speed, $\varepsilon_{wind,\,i} = N(0,\, b{\cdot}w_i)$. Synthetic data is partitioned based on true wind speed bands as presented in Table III. Therefore adding the noise ε_{wind} with variance linearly increasing with wind speed, results in partitions simulating varying data uncertainty as we move from the in-domain to out-of-domain ones. Such design aims to capture the empirical observation that the most severe wind conditions encountered in the dataset are the most uncertain.

where i = 1, …, M stands for the i-th record, w is the true wind speed, $a = 0.025$ (at power 40 MW the standard deviation of heteroscedastic power noise is 1MW) and $b = 25$ (at wind speed 40 kn the standard deviation of heteroscedastic wind noise is 1MW). The synthetic power with noise is defined as:

$$y'_i = y_i + \varepsilon_{power,\,i} + \varepsilon_{wind,\,i}$$

4 Evaluation metrics and results

4.1 Shift metrics

Accurately assessing uncertainty estimation and robustness to distributional shift is a major objective of contemporary ML research *(Malinin et al. 2021)*. Robustness to distributional shift is usually defined as the ability of the model to preserve equally good performance when tested with a shifted dataset, or in other words a dataset that has been generated from a different process. Empirically, robustness is usually assessed by comparing the predictive performance of multiple models on different datasets, one of which is considered to match the original data distribution. The model with the smaller degradation in performance is prefered.

Uncertainty estimation is the ability of the model to provide a quantitative number that represents the model's confidence along with each prediction. It is often expressed in probabilistic terms, with the value of the prediction as the mean of a distribution, and the uncertainty captured with some measure of dispersion (e.g. the variance for Gaussian distributions). Evaluating uncertainty estimation is a challenging task mainly because there is no direct way to evaluate the performance of the models, as there is no "ground truth" for uncertainty scores. Usually uncertainty estimation is assessed by the ability of the model to identify artificially shifted data points. Moreover, there is also an effort to jointly assess robustness to distributional shifts and uncertainty: if a model cannot provide accurate predictions due to distributional shift, it should at least provide high uncertainty estimates. To jointly assess robustness and uncertainty estimation *A. Malinin et al. (2021)* introduce the area under error-retention curves (Fig.3), for the R-AUC retention curve and the F1-AUC retention curve.

The key idea behind error-retention curves for a given error metric is calculating the metric at increasing fractions of the dataset, starting with data points with the least uncertainty, where the model's accuracy is expected to be the highest, and progressively adding points until the whole dataset is considered. For the part of the dataset that is excluded, optimal performance is assumed. For an MSE retention curve *(Lakshminarayanan et al. 2017; Malinin 2019)* for example, every point (x,y) on the retention curve represents the MSE, as calculated for the x (x is between 0-1) fraction of the dataset with the lowest uncertainty, and assuming the error at the rest of the dataset is zero. For x=1 we recover the standard MSE over the whole dataset. The Area Under this Curve (AUC) takes into account both the accuracy of the model (lower MSE leads to smaller AUC) and the correlation between uncertainty and error (stronger correlation leads to smaller AUC). This metric, the area under the MSE retention curve, is referred to as R-AUC. *Malinin et al. 2021* also propose considering the corresponding metric for the F1 score, namely the F1-AUC, because the R-AUC can be influenced disproportionately by the accuracy of the predictions in comparison to the accuracy of the uncertainty estimates. For the F1 score retention curve, in contrast to the MSE retention curve, the data points are sorted in descending order of uncertainty, and for every retention fraction the most uncertain part of the dataset is considered (see *Malinin et al. 2021*). Figure 3 depicts two illustrative examples of retention curves, one using the MSE as the error metric and one using the F1 score. Please note that for the R-AUC a smaller area is better, as smaller MSE is better, while for the F1-AUC on the other hand a larger area is better, since larger F1-score values are better. The two plots refer to models with equal MSE and F1-score respectively, isolating the effect of varying degrees of correlation between uncertainty and error on the AUC metrics. The orange curve corresponds to the error ranking based on the uncertainty estimates of the model evaluated. The blue curve represents the worst case scenario, where the data points are considered in random order, which is the case when the uncertainty estimates and the prediction errors are uncorrelated. The green curve represents the case of perfect correlation of error and uncertainty estimates (optimal scenario).

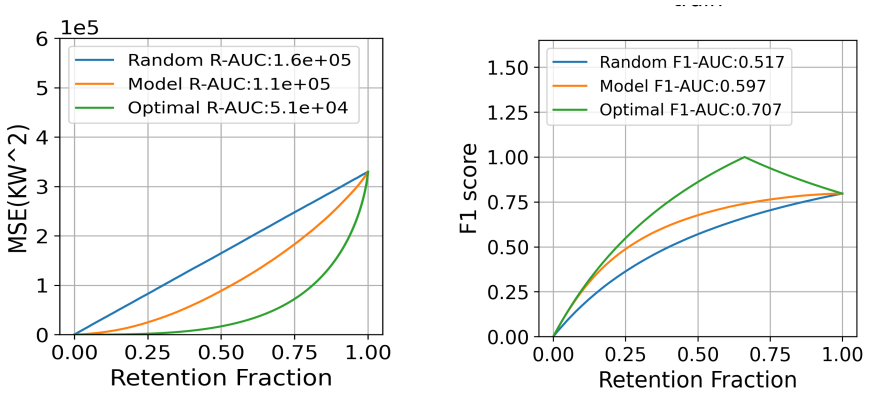


Fig.3 : Representative examples of MSE and F1 retention curves.

4.2 Experimental results

To evaluate the proposed dataset partitioning through the prism of uncertainty, we use two baseline models, able to capture both epistemic and aleatoric uncertainty, in the form of an ensemble. These are: a) an ensemble of 10 variational inference neural networks (VIs) and b) an ensemble of 10 deep neural networks (DNNs) (*Duerr et al. (2020)*). Each model predicts the parameters of the conditional Normal distribution $N(\mu(\mathbf{x}_i), \sigma(\mathbf{x}_i))$ of the target given $\mathbf{x}_i$. The variance of $\mu(\mathbf{x}_i)$ across the members of the ensemble corresponds to the epistemic uncertainty and the mean of $\sigma^2(\mathbf{x}_i)$ across the members is a measure of aleatoric uncertainty (*Malinin et al. (2021)*). For both methods we use the same architecture: 2 hidden layers with 50 and 20 nodes and softplus activation function. The output layer has 2 nodes and a linear activation function. To satisfy the constraint of positive standard deviation the second output is fed through a softplus function and a constant 10^{-6} is added for numerical stability as proposed by *Lakshminarayanan et al. (2017)*. For optimization, we use the negative log likelihood loss function and the Adam optimizer with a learning rate of 10^{-4}. The number of epochs is defined by early stopping, with patience set to 20 epochs monitoring the mean absolute error (MAE) of the dev_in set. These two baselines are standard methods for the estimation of the conditional distribution that describes the target and are both reported for completeness, as no significant differences are expected taking into account that they have similar structure.

Power-speed simulations

For qualitative model evaluation, simulations of vessel performance in relation to weather and/or operational conditions are used. The data generation process (synthetic model) of the proposed evaluation protocol, offers a major advantage, allowing for direct comparison between model estimations and the ground truth solution. The variance of the generated synthetic data on the respective conditions, expressing the aleatoric uncertainty, is due to injected noise and is directly compared with the model's predictive aleatoric uncertainty.

Power-speed simulations produced by the baseline ensemble of DNNs for the design draft state and varying true wind conditions are illustrated in Fig.4 . The first 3 rows (0 kn, Head 16kn, Tail 16kn) correspond to in-domain wind conditions and the last 2 rows (Head 33kn, Tail 33kn) to out-of-domain conditions. For the in-domain simulations, the estimated average trend is in good agreement with the ground truth solution within total estimated variance. Out-of-domain simulations exhibit a relatively pronounced underestimation of the power-speed trend at high speeds that is not explained by the estimated uncertainty in the case of head wind with speed 33 kn. Another important observation is that the estimated aleatoric uncertainty closely follows the pattern of the real target noise (depicted as blue data points in plots) for all simulated wind conditions. This is a strong indication that the estimated aleatoric uncertainty is well calibrated. Although an incremental tendency is observed at the extrapolated region of high vessel speeds, epistemic uncertainty is not considerable for the in-domain simulations (speeds that exceed the maximum speed recorded in the training data). An interesting out of domain observation is for tail wind with speed 33 kn, where notable epistemic uncertainty is found for both small and high vessel speeds.

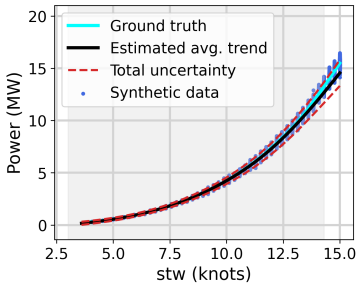
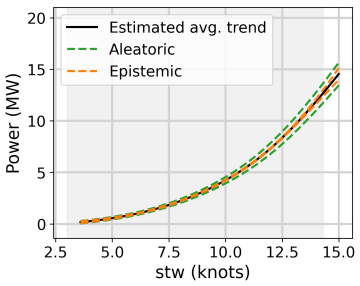
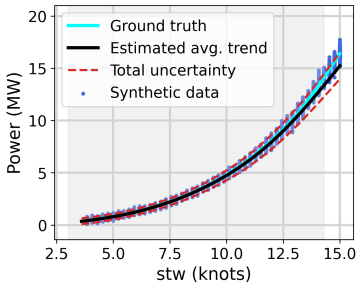
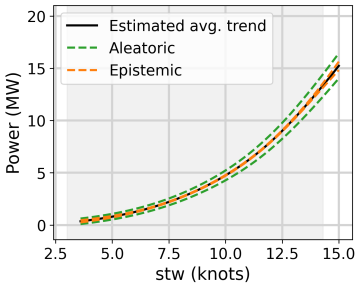
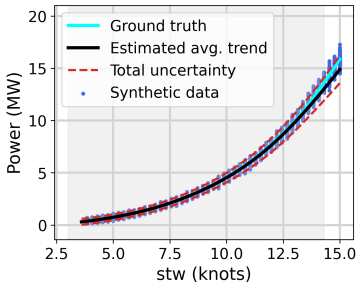
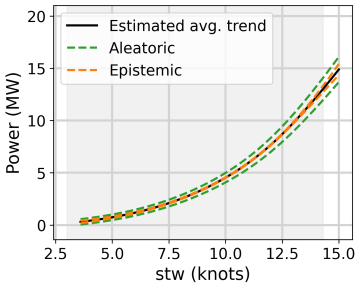
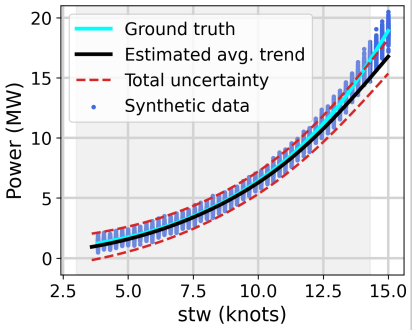
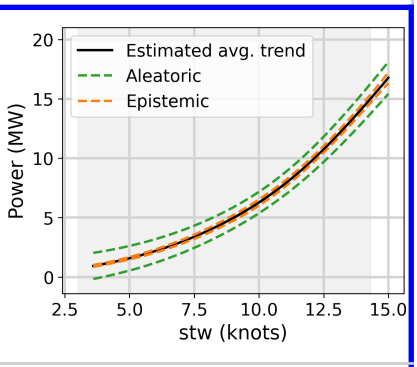
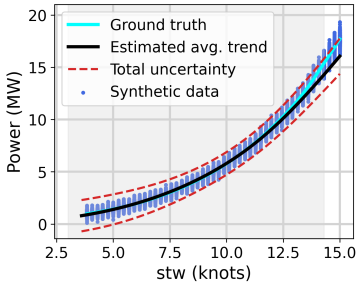
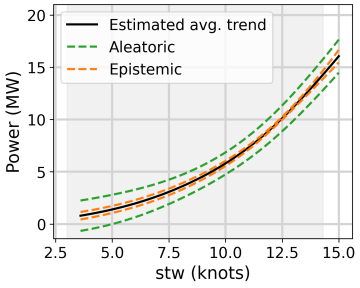
True wind conditions	Simulations	
0 kn		
Head 16 kn		
Tail 16 kn		
Head 33 kn		
Tail 33 kn		

Fig.4 : Power-speed simulations for the design draft state and various true wind conditions estimated by the baseline ensemble of DNNs. In the first column, the average estimated trend (black solid line) and the estimated total uncertainty (red dashed lines) is compared with the ground truth solution (cyan solid line). Synthetic data (blue points) are the noisy data produced by the synthetic model (generator) on the respective simulation conditions. Synthetic data indicate the real data spread due to target noise

that should be captured by the estimated aleatoric uncertainty. The second column of plots depicts the estimated aleatoric (green dashed lines) and epistemic (orange dashed lines) uncertainty along with the estimated average power-speed trend. All uncertainty boundaries correspond to ±3 standard deviations. The grey box in the background of the graphs delimits the speed range of training data regardless of the other feature dimensions. The blue box denotes out-of-domain simulations with respect to wind conditions.

Classical Metrics

For the evaluation of the robustness of models' performance to distributional shifts, we use the RMSE and MAE scores. In Table V, the predictive performance of the two baseline methods, (an ensemble of DNNs and an ensemble of VIs) is presented. It is observed that both methods exhibit the same trends; They have similar scores for the in-domain partitions, while the models' performance deteriorates for the out-of-domain partitions, having greater errors the more shifted the partition is. It is found that eval_out is the most challenging out-of-domain set, as expected, taking into account the partitioning method described in subsection 3.2.3.

Table V : Predictive performance of in-domain and out-of-domain canonical partitions of the synthetic set.

Data	RMSE (kW)		MAE (kW)	
	Ens. DNN	Ens. VI	Ens. DNN	Ens. VI
Dev in	572	571	436	436
Eval in	574	573	437	436
Dev out	691	703	547	555
Eval out	732	733	574	574

Uncertainty Metrics

Mean square error (MSE) and F1 retention performance metrics (R-AUC and F1-AUC respectively) for two baselines under study are presented in Table VI and the respective retention curves for the ensemble of VIs are illustrated in Fig.5 and Fig.6. Following the methodology proposed by *Malinin et al. (2021)* we use the MSE as the error metric and for F1 scores we consider acceptable predictions those with $MSE < (500\,kW)^2$. As the uncertainty measure, we use the total variance (i.e. due to aleatoric and epistemic uncertainty). A good model should have a small R-AUC and large F1-AUC. The R-AUC metric is comparable for the in-domain partitions, as expected for data that are drawn from the same distribution. Out-of-domain partitions have larger R-AUC indicating that model performance, in terms of robustness and/or uncertainty estimation, degrades when shifting away from the training convex hull, with eval_out being the most challenging set to be modelled, in accordance with the fact that eval_out is the most shifted dataset. Similar trends are observed in F1 retention curves; in-domain partitions result in similar F1-AUC values while F1-AUC decreases for the out-of-domain partitions.

Table VI : Retention performance for in-domain and out-of-domain canonical partitions. F1 scores are defined considering an upper threshold $MSE = (500\ kW)^2$ for the acceptable prediction errors.

Data	R-AUC		F1-AUC		F1 @ 95%	
	Ens. DNN	Ens. VI	Ens. DNN	Ens. VI	Ens. DNN	Ens. VI
Dev in	112345	112109	0.595	0.596	0.791	0.791
Eval in	111027	110829	0.597	0.597	0.793	0.793
Dev out	206090	210019	0.499	0.498	0.690	0.685
Eval out	218342	217173	0.505	0.505	0.685	0.684

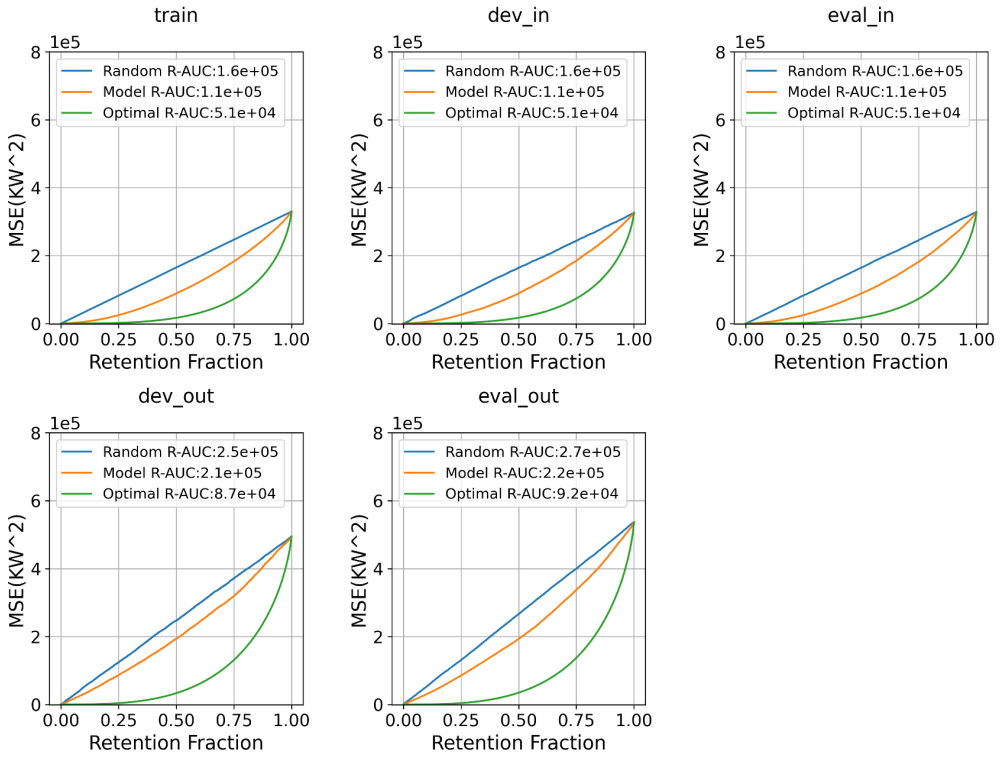


Fig.5 : MSE retention curves (R-AUC) of the ensemble of VIs for the canonical partitions of synthetic set. The orange curve is the retention curve of the ensemble. The blue curve represents the worst case scenario and the green curve the optimal scenario. The R-AUC model score (reported at the legend) is comparable for the in-domain partitions and increases the more shifted the dataset is.

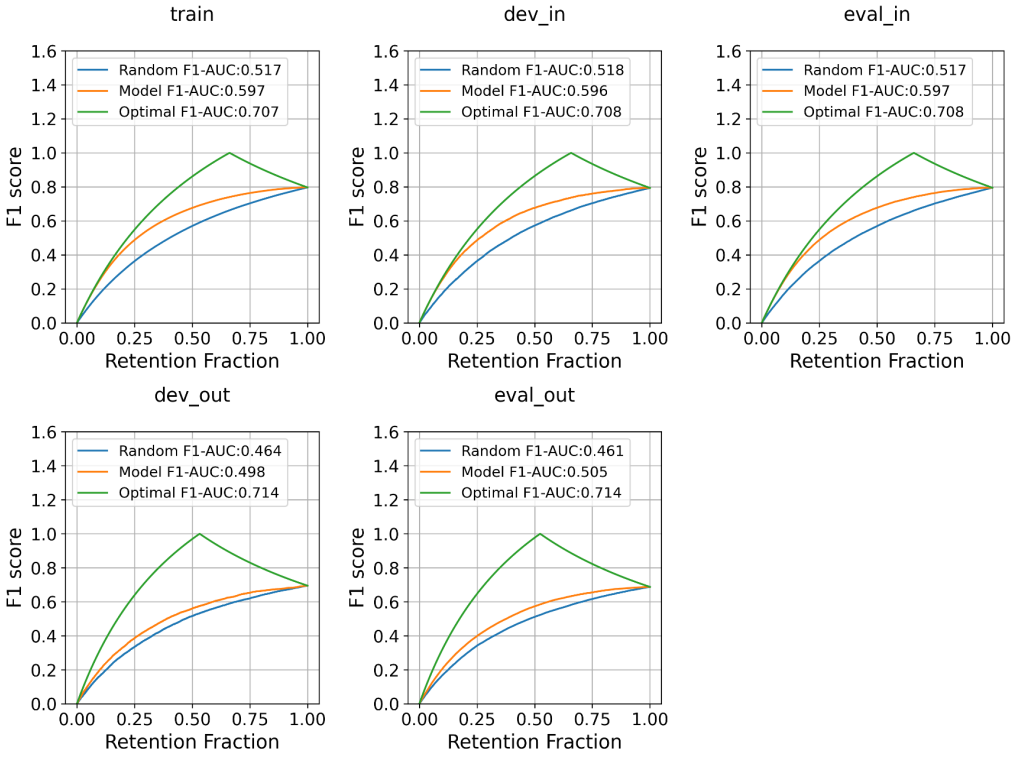


Fig.6 : F1 retention curves for the ensemble of VIs under the canonical partitions split. The orange curve is the retention curve of the ensemble. The blue curve represents the worst case scenario and the green curve the optimal scenario. The F1-AUC model score (reported at the legend) is similar for the in-domain partitions while out-of-domain partitions have smaller scores.

5. Conclusions

From the inception of these methods, the accuracy and real-world utility of data-driven vessel modelling has been easy to claim, and impossible to prove - prompting justified scepticism of this approach. In this work, we presented an evaluation methodology based on a well-considered dataset splitting scheme that aims to reveal models' deficiencies to substantial distributional shifts. We examine the results through the lens of predicted uncertainty as part of the process, as this offers useful information about the model's fitness when dealing with uncertain and noisy regions in the modelled dataset. In overall, we find that splitting the dataset in the proposed manner successfully exposes models' performance drop when moving from in-domain to out-of-domain dataset splits, as demonstrated by the classical metrics. More importantly, we showed with two baseline models that predictive uncertainty correlates well with such drops, making it possible to assess the model's performance after deployment, without access to the true target values. Main goal of this study is to encourage the shipping industry to trust and employ deep learning algorithms with predictive uncertainty in vessel performance modelling. Future research could focus on inherent selection biases in the dataset sampling process and how they might affect the model's generalisation ability.

Acknowledgements:

The team would like to express their thanks to the technical department of Laskaridis Shipping for allowing us to use their data and for the close cooperation in the process of building the synthetic model for their vessel, by sharing design related data and sea trials results.

References

ALIPANAHI, B.; DELONG, A.; WEIRAUCH, M. T.; FREY, B. J. (2015), *Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning,* Nature biotechnology, 33(8):831–838.

AMODEI, D; OLAH, C.; STEINHARDT, J.; CHRISTIANO, P.; SCHULMAN, J; MANE, D. (2016), *Concrete problems in AI safety,* arXiv preprint arXiv:1606.06565.

BERTRAM, V. (2012), *Practical Ship Hydrodynamics,* Butterworth & Heinemann, Oxford.

BOS M. (2018), *An Ensemble Prediction of Added Wave Resistance to Identify the Effect of Spread of Wave Conditions on Ship Performance,* 3rd HullPIC Conf., Redworth.

BOSE, N. (2008), *Marine Powering Predictions and Propulsors,* The Society of Naval Architects and Marine Engineers. New York.

CARLTON, J.S. (2018), *Marine Propellers and Propulsion,* 4th Edition. ButterworthHeinemann, Oxford, UK.

CORADDU, A.; ONETO, L; BALDI, F; CIPOLLINI, F; ATLAR, M; SAVIO, S (2019), *Data-Driven Ship Digital Twin for Estimating the Speed Loss caused by the Marine Fouling,* Ocean Engineering

DER KIUREGHIAN, A.; DITLEVSEN, O. (2009), *Aleatory or epistemic? Does it matter?,* Structural safety 31.2: 105-112.

DUERR, O.; SICK, B.; MURINA, E. (2020), *Probabilistic Deep Learning: With Python, Keras and TensorFlow Probability*, Manning Publications, ISBN:9781617296079.

FARKAS, A; DEGIULI, N; MARTIC, I; DEJHALLA, R (2020), *Impact of Hard Fouling on the Ship Performance of Different Ship Forms,* Journal of Marine Science and Engineering 2020, 8, 748; doi:10.3390/jmse8100748.

FUJIWARA, T; UENO, M; IKEDA, Y. (2006), *Cruising performance of a large passenger ship in heavy sea,* Proc. of Sixteenth International Offshore and Polar Engineering Conference, Vol. III.

GAL, Y. (2016), *Uncertainty in Deep Learning,* Ph.D. thesis, University of Cambridge.

GOAN, E.; FOOKES C. (2020), *Bayesian neural networks: An introduction and survey,* Case Studies in Applied Bayesian Data Science. Springer, Cham, 2020. 45-87.

GÓRSKI, W; MICHNIEWICZ, J; SZLENDAK, A. (2021), *Using Unsupervised Machine Learning for Building Ship Performance Reference Model,* 6th HullPIC Conf., Pontignano.

GONZALEZ, C; ARANGO, D. L. (2019), *Techniques for the Automated Detection of Anomalies and Assessment of Quality in High-Frequency Data Collection Systems,* 4th HullPIC Conf., Gubbio.

HARDT, M.; RECHT, B. (2021), *Patterns, predictions, and actions: A story about machine learning,* mlstory.org, arXiv preprint arXiv:2102.05242.

HINTON, G.; DENG, L.; YU, D.; DAHL, G. E.; MOHAMED, A.-R.; JAITLY, N.; SENIOR, A.; VANHOUCKE, V.; NGUYEN, P.; SAINATH, T. N.; KINGSBURY, B. (2012), *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, Signal Processing Magazine, IEEE, 29(6): 82–97.

HOLTROP, J.; MENNEN, G.G.J. (1982), *An approximate power prediction method*, International Shipbuilding Progress. 29. 166-170.

ISO (2015), *Ships and marine technology – Guidelines for the assessment of speed and power performance by analysis of speed trial data,* ISO 15016:2015.

ISO19030-2 (2016), *Ships and marine technology - Measurement of changes in hull and propeller Performance - Part 2: Default method,* ISO, Geneva.

JOURNÉE, J.M.J.; MEIJERS, J.H.C. (1980), *Ship routeing for optimum performance,* TU Delft.

KRIZHEVSKY, A.; SUTSKEVER, I; HINTON, G. E. (2012), *Imagenet classification with deep convolutional neural networks,* Advances in Neural Information Processing Systems.

LAKSHMINARAYANAN, B.; PRITZEL, A; BLUNDELL C. (2017), *Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles,* Advances in Neural Information Processing Systems 30.

LEVANTIS, M.; PAERELI, S.; ENSTRÖM, A; BETZ, W. (2020), *Speed-Power Models – A Bayesian Approach,* 5Th HullPIC Conf, Hamburg.

MALININ, A. (2019) *Uncertainty Estimation in Deep Learning with application to Spoken Language Assessment,* Ph.D. thesis, University of Cambridge.

MALININ, A; BAND, N; CHESNOKOV, G; GAL, Y; GALES, M. JF.; NOSKOV, A; PLOSKONOSOV A; PROKHORENKOVA, L.; PROVILKOV, I; RAINA, V.; RAINA, V; ROGINSKIY, D; SHMATOVA, M; TIGAS, P; YANGEL, B. (2021), *Shifts: A dataset of real distributional shift across multiple large-scale tasks,* arXiv preprint arXiv:2107.07455.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN J. (2013), *Efficient estimation of word representations in vector space,* arXiv preprint arXiv:1301.3781.

MOLLAND, A.; TURNOCK, S.; HUDSON, D. (2017) *Ship Resistance and Propulsion,* Cambridge University Press.

MORENO-TORRES, J. G.; RAEDER T.; ALAIZ-RODRIGUEZ, R.; CHAWLA, N. V.; HERRERA, F. (2012), *A unifying view on dataset shift in classification*, Pattern recognition 45.1: 521-530.

NIKOLOPOULOS, L.; BOULOUGOURIS E. (2018), *A Study on the Statistical Calibration of the Holtrop and Mennen Approximate Power Prediction Method for Full Hull Form, Low Froude Number Vessels,* Journal of Ship Production and Design. 35. 10.5957/JSPD.170034.

PARK, J.; KIM, B.; SHIM, H.; AHN, K.; PARK, J. H.; JEONG, D.; JEONG, S. (2018), *Hull and Propeller Fouling Decomposition and Its Prediction based on Machine Learning Approach,* 3rd HullPIC Conf., Redworth.

RAMSUNDAR, B.; KEARNES, S.; RILEY, P.; WEBSTER, D.; KONERDING, D.; PANDE, V. (2015), *Massively multitask networks for drug discovery,* arXiv preprint arXiv:1502.02072.

SEO; K.-C.; ATLAR, M.; GOO, B. (2016), *A Study on the Hydrodynamic Effect of Biofouling on Marine Propeller,* Journal of the Korean Society of Marine Environment & Safety Vol. 22, No. 1, pp. 123-128.

TOWNSIN, R.L.; BYRNE D; SVENSEN T.E.; MILNE, A. (1981), *Estimating the technical and economic penalties of hull and propeller roughness,* Trans SNAME.

TSARSITALIDIS, V.; ROSSOPOULOS, S. (2018), *ISO 19030 - The Good, the Bad and the Ugly,* 3rd HullPIC Conf., Redworth.

TSUJIMOTO, M.; SHIBATA, K.; KURODA, M.; TAKAGI K. (2008), *A Practical Correction Method for Added Resistance in Waves,* J. JASNAOE, Vol.8.

VAN LAMMEREN W.P.A.; VAN MANEN, J.D.; OOSTERVELD, M.W.C. (1969), *The Wageningen B-Screw Series,* Transactions of the Society of Naval Architects and Marine Engineers, Vol. 77, pp. 269–317.

WANG, H.; YEUNG, D.Y. (2020), *A survey on Bayesian deep learning,* ACM Computing Surveys (CSUR) 53.5 (2020): 1-37.

ZHOU, J; TROYANSKAYA O. G. (2015), *Predicting effects of noncoding variants with deep learning-based sequence model,* Nature methods, 12(10):931–934.