Enhancing early-stage energy consumption predictions using dynamic operational voyage data

A grey-box modelling investigation

Odendaal, Kirsten; Alkemade, Aaron; Kana, Austin A.

DOI
10.1016/j.ijnaoe.2022.100484

Publication date
2023

Document Version
Final published version

Published in
International Journal of Naval Architecture and Ocean Engineering

International Journal of Naval Architecture and Ocean Engineering 15 (2023) 100484

Contents lists available at ScienceDirect

International Journal of Naval Architecture and Ocean Engineering

journal homepage: http://www.journals.elsevier.com/international-journal-of-naval-architecture-and-ocean-engineering/

Enhancing early-stage energy consumption predictions using dynamic operational voyage data: A grey-box modelling investigation

Kirsten Odendaal [a, b, *], Aaron Alkemade [b], Austin A. Kana [a, **]

[a] Department of Maritime and Transport Technology, Delft University of Technology (TU Delft), the Netherlands
[b] De Voogt Naval Architects (Feadship), Haarlem, the Netherlands

ARTICLE INFO

Article history:
Received 2 November 2021
Received in revised form
30 August 2022
Accepted 5 September 2022
Available online 14 September 2022

Keywords:
Artificial neural network
Grey-box modelling
Yachting
Propulsion and auxiliary power
Bootstrap aggregation
Ensemble methods

ABSTRACT

The adverse human contribution to global climate change has forced the yachting industry to acknowledge the need to reduce its environmental impact due to the client's increasing pressure and potential future regulations to limit the ecological effects. Unfortunately, current real-world data presents a significant disparity between predicted and actual gathered energy consumption results. Thus, this research aims to develop an approach to accurately predict total dynamic Energy Consumption (EC) using real operation voyage data for the improved early-stage design of future yachts. A Grey-Box Modelling (GBM) solution combines: physics-based White-Box Models (WBM); and Black-Box Model (BBM) artificial neural networks to provide estimations with high accuracy and improved extrapolation capacity. The study utilizes ten months of onboard continuous monitoring data, hindcasted weather, and voyage information from a *Feadship* fleet yacht. Upon applying a sequential modelling methodology, predictions are compared between the three model categories, indicating propulsion and auxiliary estimates fall within 3% and 7% error of operational conditions. The study is then continued using external range datasets to evaluate the extrapolation potential. While GBM improvements are seen over the BBM, limitations were directly related to the strength between dynamic WBM input-output correlations. Ultimately, GBM's have the potential to improve both accuracy and extrapolation ability over existing WBM and BBM's; however, much is dependent on the strength of the input-output relationships.

1. Introduction

The adverse human contribution to global climate change has been recognized as a significant risk to future generations. As such, the shifting perspective towards sustainability is currently being driven by both public image and social responsibility. Therefore, to meet these worldwide demands, the yachting industry has acknowledged the need to reduce its environmental impact due to the increasing pressure of consumers and future regulations to limit the effects on the environment.

However, a hurdle is currently being faced within the maritime sector regarding observed discrepancies between real-operation voyage data and predicted energy consumption results. These gaps have led to much discussion on future sustainable yacht design implications. Thus, the objective of this paper is to,

> *Develop an approach to accurately predict total dynamic Energy Consumption (EC) using real operation voyage data for the improved early-stage design of yachts*

In order to address and evaluate whether the proposed solution thoroughly meets the research objective, a series of method requirements have been established.

1. Estimate power for propulsion and auxiliary systems under dynamic conditions within ±15% error with 95% Confidence Intervals
2. Ability to proportion both auxiliary and propulsion power consumption independently
3. Be based on a modular methodology to easily incorporate various estimation tools and results

* Corresponding author. Department of Maritime and Transport Technology, Delft University of Technology (TU Delft), the Netherlands
** Corresponding author.
E-mail addresses: kirst.odendaal@gmail.com (K. Odendaal), A.A.Kana@tudelft.nl (A.A. Kana).

Peer review under responsibility of The Society of Naval Architects of Korea.

https://doi.org/10.1016/j.ijnaoe.2022.100484

Nomenclature

α_{360}	Relative Wave Direction
α_{sun}	Sunlight Factor
β_{360}	Relative Wind Direction
Ψ	Ship Heading
ADAM	Adaptive Momentum Estimation
BBM	Black-box Model
E_e	Surface Radiation
GBM	Grey-box Model
H_s	Significant Wave Height
HCI	Hull Cleaning Interval
P_a	Target - Auxiliary Power
P_{hvac}	WBM - HVAC Power
$P_{s,cw}$	WBM - Calm-water Shaft Power (H&M model)
$P_{s,t}$	WBM - Total Shaft Power (H&M + Wind + Wave model)
P_s	Target - Shaft Propulsion Power
P_{total}	Target - Total Power
PC	Percent Coverage
RH	Relative Humidity
S	Sailing Factor
SGD	Steepest Gradient Descent
T_0	Significant Wave Period
T_{air}	Air Temperature
T_{sea}	Sea Temperature
V_s	Speed over Ground
V_{wi}	Wind Speed
WBM	White-box Model

4. Be able to deal with discrepancies and errors in voyage report data

A minimum error was established as a threshold to be used as a comparative baseline metric to validate the approach accuracy. During the early design stages of ship design, low-fidelity tools are commonly applied. However, it is known that such approaches typically have lower accuracy in favour of lower computational demand. Thus, after careful industry, academic, and literature deliberation, a minimal function threshold of ±15% is deemed an appropriate starting point for the early design stage focus. The proposed contributions of the study are.

a) Methodology reproducibility (propulsion power estimation) and new application success (auxiliary power estimation and superyacht case vessel) within the maritime industry.
b) A detailed comparison investigation that explores the global performance of various modelling categories for both interior and exterior data region prediction capabilities, including the feasibility of GBM modelling aggregation.
c) Adaptation of a statistical uncertainty quantification technique to improve interpretability and confidence for maritime applications.

2. Literature investigation

Modelling of physical systems is usually applied by implementing three mathematical approaches: White-Box Models (WBM), Black-Box Models (BBM), or Grey-Box Models (GBM).

As outlined by Zwart (2020), Coraddu et al. (2018), Leifsson et al. (2008), the White-box approach models a physical system entirely using physical laws and deterministic first-principle relations, which is based on prior knowledge. The Black-box approach models a system entirely based on observed data, such as input-output measurements, and requires no prior knowledge of the overall system. These methods usually focus on a range of statistical approaches, such as auto-regressive models or machine learning methods. Well trained BBMs can be more accurate than WBMs. However, BBMs require large amounts of high-quality data for model training, and more importantly, lack interpretability and extrapolating ability in contrast to WBMs. As further highlighted by Leifsson et al. (2008), it is also possible to have models that deeply integrate both the White- and Black-box approaches, generating what is known as a Grey-Box Model (GBM). Ultimately, a GBM attempts to combine both WBM and BBM advantages to overcome both individual solution's apparent drawbacks.

2.1. White-box modelling

WBM techniques are universally used within the early design stages of all marine sectors. General WBMs can decompose complex problems, such as propulsive and auxiliary demand, into smaller sub-models for increased physical insight. Usually, ship propulsion power is determined as a function of total resistance, which can be decomposed into two main components: calm-water resistance (Holtrop and Mennen (1982)), and added resistance due to dynamic contributions related to wind, waves, currents, and fouling (MAN Diesel & Turbo (MAN and Turbo, 2006)). Additionally, auxiliary systems are typically composed of multiple sub-models: Hotel loads such as HVAC power, chiller power, water supply, lighting, etc., and rudder heading control and stabilization power (Boertz (2020)). Unfortunately, this creates scenarios where many individual parameters are required to suitably solve such problems. Additionally, due to the complex nature of dynamic interactions, such models, while offering a high degree of practicality, lack the necessary accuracy required for later design stage calculations. While CFD, model tests, and electric load simulations can be used to offer a much more robust and accurate solution, these techniques are not only computationally expensive, they require a high degree of a priori system information generally not available within the initial design stages of yacht design. As such, many implemented solutions can be considered semi-empirical and not true a WBM. Nonetheless, due to the relative practicality and generally low fidelity of such models, they are commonly viewed as WBMs, (Yang et al. (2019)).

2.2. Black-box modelling

On the other hand, BBM techniques have been successfully applied within the maritime industry to evaluate dynamic effects accurately. Numerous studies investigating the potential of propulsion power estimation (Parkes et al. (2018)), fuel consumption estimation (Bal Beşikçi et al. (Bal Beşikçi et al., 2016)), added resistance prediction studies (Cepowski (2020)), and even energy efficiency indexes (Kim et al. (2020)) have been analyzed with consistent success. Most if not all studies have typically applied either Artificial Neural Networks (ANN) or Gaussian Process Regression (GPR) machine learning techniques. ANN's are computationally quicker during training, GPR provides inherent

uncertainty characteristics, and both provide highly accurate estimations. Ultimately, it can be concluded that, while these methods are powerful, (Kim et al. (2020)) demonstrated that data quality, data amount, and the correct parameter tuning are driving factors in regard to accuracy and generalization capabilities. Unfortunately, the authors are unaware of BBM literature that has explicitly studied yachts or auxiliary power estimations within the maritime industry. However, extrapolation of the method is possible as success has been documented for various vessel classes and power estimations within building engineering (Neto and Fiorelli (2008), Kalogirou and Bojic (2000)).

2.3. Grey-box modelling

GBM is a state-of-the-art method that combines both the WBM and BBM to overcome each individual's deficiency. Successful implementation has been observed by Zwart (2020), Coraddu et al. (2018), and Leifsson et al. (2008). Unfortunately, only a few literature sources are available for study due to the method's relatively new introduction within the maritime field. Much like the BBM, the GBM application within both the yachting industry and auxiliary power evaluations has yet to be thoroughly investigated. Nonetheless, the approach shows a great deal of promise, where a detailed literature comparison of the various modelling approaches has been developed and detailed by Yang et al. (2019). It was concluded that the GBM approach has more accuracy than WBMs, requires less historical data than BBMs, has good model interpretability, and improves the extrapolation capacity; thus, avoiding unreasonable results to a certain degree.

This research aims to predict and proportion the dynamic powering components within a yacht quickly and accurately for early-stage design. Based on the above-detailed comparison, the approach most likely to satisfy all established method requirements is a grey-box modelling solution using artificial neural networks. The hope is that the WBM retains the physical behaviour of a vessel, whereas the BBM scales the output to fit the operational data. While, Gaussian process regression could be applied with likely success, due to the successful use within the maritime industry, high data-scaling potential, and the general ease of implementation, the ANN algorithm will be used as the BBM.

Ultimately, propulsion and auxiliary models have been developed independently for each dynamic power contribution, which has then been aggregated to estimate total energy demand via a bottom-up aggregation approach. Each model is to incorporate its own WBM(s) orientated towards each load type, as shown in Fig. 1. By nature, the GBM is a modular solution that can incorporate all data types, be it empirical, semi-empirical, CFD, or data-sensor features. However, associated risks of success, such as data quantity and quality, as well as WBM accuracy, completeness, and relevance, must be acknowledged.

3. Theoretical overview

As a GBM solution was judged most suitable, both WBM and BBM model limitations and assumptions must be considered for optimal performance. Ultimately, four WBMs and one BBM solution have been investigated and outlined, respectively.

3.1. White-box overview

1. *Calm-Water Resistance Calculation (WBM Propulsion):* The calm-water calculation is based on the Holtrop and Mennen (H&M) (Holtrop and Mennen, 1982; Holtrop, 1984) approach, and as such, has two main limitations: vessel type and efficiencies. The model is limited to monohull displacement vessels with moderately slender bodies and low Froude numbers.

$$\begin{matrix} & Fr & \leq 0.45 \\ 3.9 \leq & L/B & \leq 9.5 \end{matrix}$$

Additionally, early-stage efficiencies are limited to low-order empirical formulations heavily dependent on propeller design and vessel shape. Hull efficiency, rotative efficiency, and open-water efficiencies are generally heavily dependent on the vessel's speed, whereas transmission efficiency losses are relatively low and stable. Thus, both literature sources (Holtrop and Mennen (1982)) and model-scale results are considered to validate the efficiencies within the early stages.

2. *Wind Resistance Calculation (WBM Propulsion):* The wind resistance component is a subset of the main calm-water proportion. The developed model is based on the ITTC (Seakeeping Commitee of the 29th ITTC, 2018) approach where the wind profiles are determined using Computational Fluid Dynamic (CFD) wind tunnel investigations of multiple yachts. These profiles have been verified and validated with corresponding model wind tunnel tests, demonstrating a marginal powering contribution. Nevertheless, the developed model is limited to the corresponding transverse and longitudinal area profiles,

$$\begin{matrix} 158\,m^2 \leq & A_x & \leq 238\,m^2 \\ 625\,m^2 \leq & A_y & \leq 1065\,m^2 \end{matrix}$$

3. *Added Thrust in Waves Calculation (WBM Propulsion):* The wave model component is based on the SPAWAVE model (Grin (2014)), which was developed on model-scale tests, and operational comparisons. This model is limited to regressions based on waterline lengths of,

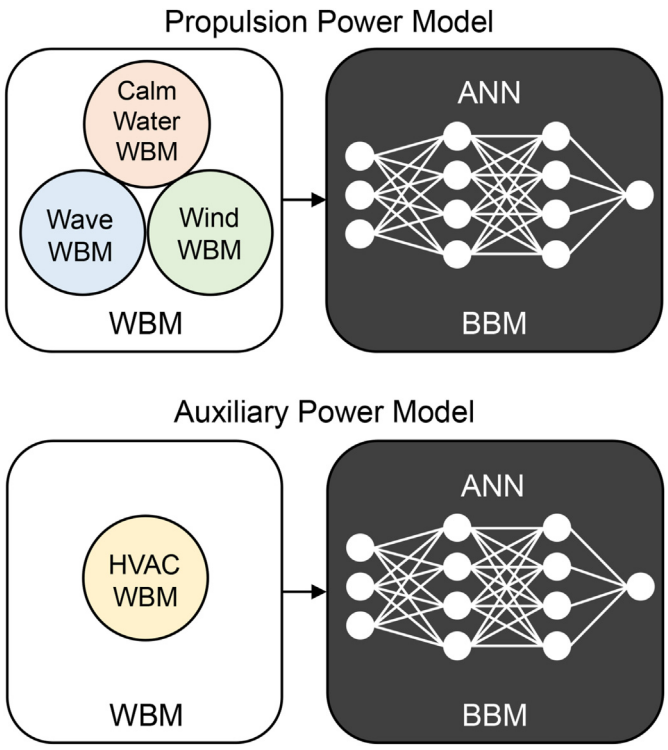

Fig. 1. ANN-GBM using a total of 3 WBM for propulsion estimation (top) and a ANN-GBM using 1 WBM for auxiliary power estimation (bottom).

$51\,m \leq L_{WL} < 108\,m$.

While the model has been verified and validated for various ship sizes, bow shapes, and wave directions, only a single-load (design draft) condition is considered. Nevertheless, this model's advantage compared to similar alternative approaches is the ability to approximate loads for all wave directions, as opposed to being restricted to only head waves.

4. *HVAC Power Calculation (WBM Auxiliary):* The heating, ventilation, and air-conditioning (HVAC) power model is based on a heat balance approach developed by Stapersma and Klein Houd (Stapersma and Klein Houd, 2012). Additionally, ASHRAE/ISO codes and standards (ASHRAE, 2013; ISO 7547, 2002) are used to assign the associated area classification comfortability requirements. However, the model does not take into account any fitting losses and personnel movements. Instead, room-by-room evaluations with maximum persons distributed per room classifications are used to evaluate required powering demand.

While propulsion white-box models can be aggregated to estimate the dynamic propulsion loading conditions, the associated HVAC model only considers a portion of the total auxiliary load. Akershoek (2018) identified that auxiliary power predictions within yachts could significantly impact total energy predictions, where expected proportions can sometimes be upwards of 50% the total demand. Thus, while potentially significant, the remaining portion will fall towards the BBM to scale the outcome results appropriately.

3.2. Black-box overview

Artificial neural networks are a computing system that is vaguely inspired by the biological networks found within the brain. These systems are composed of multiple individual synaptic components known as perceptrons. The general topology can be seen in Fig. 2 where the general input-output relationships can be described as,

$$\hat{Y}_j = g\bullet\left[\sum_{n_j=1}^{j}\left(\bar{X}^T\bullet\bar{W}\right) - b\right] \tag{1}$$

where the associated jth neuron output $\hat{Y}_j$ is a function of multiple ith feature input parameters, $\bar{X}^T = (x_0, \ldots, x_i)^T$. The corresponding weight input from the ith input and for the jth neuron is $\bar{W} = (W_{j,0}, \ldots, W_{j,i})$. The b parameter represents the associative layer bias or threshold. The activation function, $g(\bullet)$, is some continuous or discontinuous function that maps the real numbers to an interval between [-1,1] or [0,1]. As detailed by da Silva et al. (da Silva et al., 2011), when multiple perceptrons are placed in connection with one another, a Multi-Layer Perceptron (MLP) network is formed, commonly known as a Feed-Forward Neural Network (FFNN). These networks feature at least one intermediate (hidden) layer, which is placed between the input and output layers. The general architecture of an FFNN can be seen in Fig. 2.

Ultimately, many different types of neural networks exist; however, three general BBM steps should be identified for optimal performance and interpretability as.

1. *Hyperparameter Optimization:* One of the most relevant features concerning an artificial neural network is its ability to generalize the acquired knowledge, enabling the estimation of solutions using inputs that have never been encountered before. However, to achieve this state, appropriate model hyperparameters (factors specific to the network architecture) must be selected to avoid the practical problems of over- and under-fitting, (Aggarwal (2018)). This research proposes using a conventional grid-search methodology to search the hyperparameter design space for optimal solutions. While this approach can find the optimal performance region, care must be taken to keep the search permutations within reason to reduce the computational demand. Thus, the associated parameters used in the study can be seen below as,
   - Performance function: Mean Squared Error
   - Optimal layers: 1 to 3
   - Optimal neurons: 10 to 40
   - Early stopping regularization (epochs): 10 to 100
   - Dropout regularization: 0 to 0.5
   - Batch training size: 1, 32, or 64
   - Activation functions: ReLu, Sigmoid, or Tanh
   - Optimizer algorithm: ADAM or SGD
2. *Ensembling and Uncertainty Evaluation:* For improved interpretability, it is common to consider two sources of uncertainty: epistemic (σ_m) and aleatoric (σ_e). The former is related to the model's uncertainty and structure. Whereas the latter deals with the inherent error due to the natural stochasticity of the observations (Mazloumi et al. (2011)). Quantification of these sources as total uncertainty (σ_t), as shown in equation (2), can be achieved using a bootstrap aggregation (bagging) approach as introduced by Ferrario et al. (2017).

$$\sigma_t = \sqrt{\sigma_m^2 + \sigma_e^2} \tag{2}$$

Here, an ensemble of models is developed and combined using

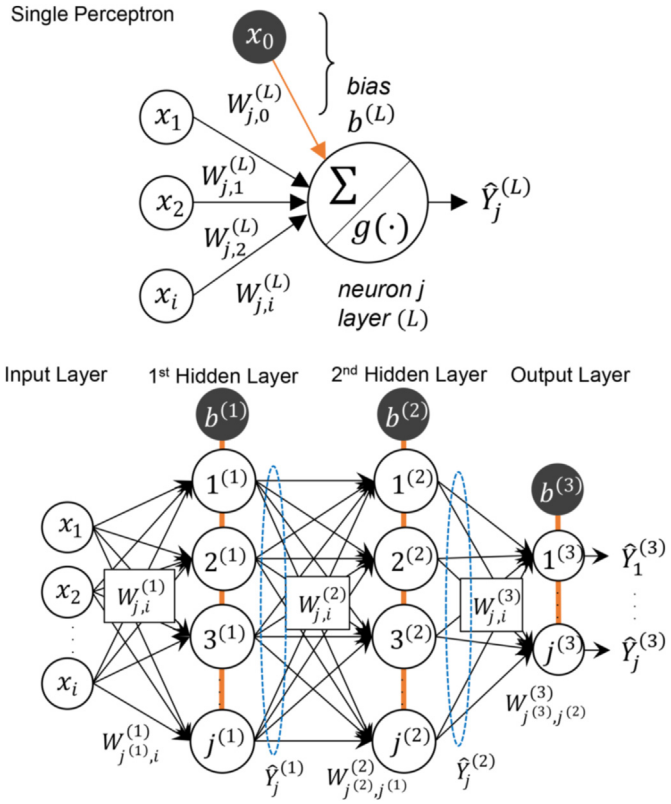

Fig. 2. A single perceptron architecture (top) and a multi-layer (3-layer) perceptron feed-forward network with weights, biases, and outputs (bottom) (da Silva et al., 2011).

datasets sampled with replacements. Ensembling inherently improves generalization as the mean performance is more robust to individual modelling irregularities. At which point, empirical probability distributions can be leveraged using general statistics to evaluate the confidence bounds using confidence level values (z) as indicated below,

$$\bar{Y} \pm z \cdot \sigma_t, \quad z = 1.96(95\%\text{Confidence}) \tag{3}$$

As detailed by Efron and Tibshirani (1993), for estimating an empirical probability distribution, the required number of ensembles (boots) will ordinarily fall within the range of 25–200. Therefore, the model building process will integrate a bootstrap approach with 30 boots to quantify total GBM uncertainty.

3. *Performance Metrics and Evaluation:* Once models are fully developed, the final performance will be evaluated by analyzing three commonly applied metrics. These include the *Coefficient of Determination* (R^2), *Mean Absolute Error (MAE)*, and the *Root Mean Squared Error (RMSE).* Ultimately, each modelling performance indicator has its advantages and disadvantages. As such, a single metric cannot thoroughly diagnose the global modelling performance. The coefficient of determination indicates the goodness of fit and does not indicate a predictive error. The MAE indicates the absolute raw error but does not account for variation. The RMSE, on the other hand, penalizes variation but can be easily skewed by a few irregular points. Thus, Botchkarev (2018) suggests that only by using multiple performance metrics can a general understanding of the modelling behaviour be successfully captured, investigated, and assessed.

4. Methodology

The total grey-box methodology can be composed of three main stages: input, modelling, and output, as identified in Fig. 3. Here, a serial (sequential) GBM approach (Zwart (2020), Leifsson et al. (2008)) is adopted, where multiple data inputs are prepared and passed through the WBM(s) as direct inputs to the BBM solution. At this point, the BBM is trained on the associated input features, which inherently alters into a GBM solution.

The approach can be further decomposed into multiple sub-steps for each sub-modelling contribution. The latter portion, which focuses on the BBM elements, has been detailed within section 3.2. Whereas the former focuses on the conversion of raw to useable input data. Here, four key steps as introduced by García et al. (2016) are identified to ensure the data is adequately prepared for successful GBM implementation.

1. *Data Integration:* Generally, data originates from multiple external sources with varying lengths and parameter dimensions. To ensure unequal data lengths are capable of merging, mutual data features must be compared and aligned, such as timestamps, latitudes, longitudes, and vessel speeds. Once alignment is verified, higher frequency datasets can be used as an interpolation foundation for the remaining in-between data entries. Such an approach intends to generate as many points as possible to develop a fuller and more diverse dataset. Unfortunately, this process is not without its flaws. Interpolating between datasets can create artificially introduced points that may not truly exist in the operational environment. Nonetheless, when many overlapping features within the multiple datasets display a high degree of parallelism, such influences are generally minor or can be eliminated through the future cleaning process. Thus, data interpolation is a practical approach to retaining as many features and data points as possible. Nevertheless, it is recommended to adopt as few varying datasets as possible to reduce the inherent interpolation errors.
2. *Noise Identification:* Noise Identification is the process of identifying and removing any outlying data entries. These irregular points have the potential to skew the output results significantly. Two approaches are commonly applied: simple parametric methods such as interquartile range (IQR) elimination and complex non-parametric techniques such as clustering or density grouping. When considering continuously monitored data from real-world operations, the law of large numbers and central limit theorem suggests that the feature distribution is expected to be normally distributed if the samples are large enough. Therefore, while a highly conservative approach, the parametric IQR is likely a suitable approach to eliminate outlier noise throughout the investigation due to the large dataset. While these methods can be compelling, they are also indiscriminate. In other words, a high understanding of the features is required as the process has the potential to reduce data ranges without intention drastically.
3. *WBM Evaluation:* Any additional missing values and irregularities not captured by the outlier detection approach are eliminated using engineering sense and self-established specifications. All data entries have been cleaned at this position within the preparation framework, and as such, no missing values shall be present within the set. Therefore, each WBM (see section 3) can be evaluated for each corresponding data entry extracted for use within the GBM framework. However, care must be taken such that all modelling assumptions and limitations are conserved.

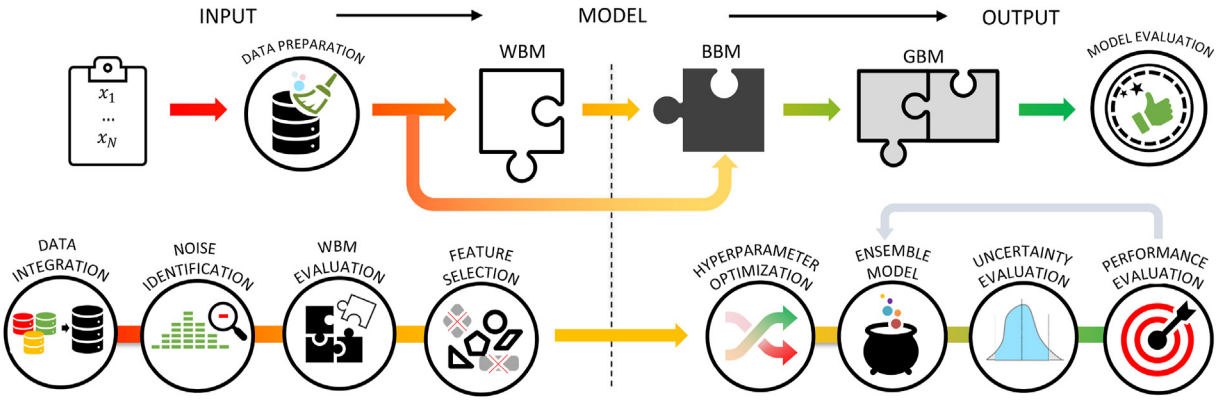

Fig. 3. Schematic of the general grey-box modelling serial methodology and each corresponding data preparation and model evaluation sub-steps.

4. *Feature Selection:* Feature Selection determines which independent features are closely related to the dependent target parameter. As detailed by Parkes et al. (2018), introducing variables that are poorly correlated to the target variables can add unnecessary complexity, negatively impacting model performance and generalization capacity. As shown by both Zwart (2020) and Parkes et al. (2018), the Spearman Correlation technique has proven effective in determining critical inputs for modelling energy consumption. The approach quantifies non-linear monotonically increasing or decreasing correlations to help guide the ANN training procedure. Unfortunately, non-monotonic dependencies such as parabolic or oscillatory relationships can not be accurately measured, as such engineering sense still needs to be applied.

Upon completing the data preparation steps, the relevant features can optimally train an ensembled artificial neural network to estimate operational energy consumption. However, depending on the powering component, propulsion or auxiliary powers, the input features vary accordingly. Thus, for each situation, a clear binary transformation input feature will be developed. The sailing operation factor can be quantified as a binary classification between *Sailing* and *Anchor* as follows,

$$S_{oper} = \begin{cases} 0 & V_s < 1\,knot \\ 1 & V_s \geq 1\,knot \end{cases} \quad (4)$$

When the vessel has a speed over ground greater than 1 knot, the vessel operation can be classified as *Sailing*. Anything less is quantified as at *Anchor*. This threshold was deemed adequate for the sake of distinguishing between pure sailing as well as low-speed manoeuvring operations. In the study's context and the general design process focus of the early-stage design, manoeuvring power demand is eliminated as this application falls outside the bounding scope.

5. Case introduction and available data

The case in question focuses on a single modern yacht vessel between the periods of 01-10-2018 and 01-07-2020. The yacht is a displacement vessel with no attached bulbous bow and a waterline length, beam, and gross tonnage of 99 m, 15 m, and 4700 g t, respectively. The main engine is a diesel-electric configuration, where the apportionment between propulsion and auxiliary power demand can be obtained directly by dedicated generator sets. These generators are linked with two electric engines and coupled via shafts towards two 5-bladed fixed-pitch propellers with a diameter of 2.5 m each. The vessel is an ideal candidate for the following case investigations for two overarching reasons.

The first is that this yacht is relatively new. As such, the implementation of onboard continuous monitoring systems allows for direct measurement of most power-consuming systems with a high-frequency 3-min sampling rate. These measurements allow for a robust and varied selection of potential feature inputs to further enhance the capabilities of the GBM. Furthermore, since the ANN uses a supervised learning approach, accurate target parameters such as individual power demand found on both main engines and generator sets for propulsion and auxiliary power are critical.

The second reason is that both the data quantity and quality are expected to be high. While the yacht is a newer vessel, 2+ years of high-frequency onboard sensor information and voyage AIS location data are available since launch. Therefore, the overall coverage of geographic locations and recorded weather conditions is quite broad. In addition to these high quantity datasets, large amounts of resources have been invested in ensuring the obtained information's quality is accurate via multi-set validations and comparisons.

The available data is obtained from three primary sources: onboard powering systems (3-min), hindcasted weather data (hourly), and ship specification/maintenance reports. The collected information over the 10-month operational range totals approximately 61,944 data points. The available variables for the model development are highlighted within Fig. 4. Based on previous literature investigations of similar studies ((Zwart, 2020; Parkes et al., 2018; Kalogirou and Bojic, 2000; Neto and Fiorelli, 2008)), it can be seen that the available data almost wholly aligns with the literature ANN input features.

6. Case evaluation

6.1. Data Integration and cleaning

Using the method described in section 4, the three primary data sources are first integrated and aligned using timestamps, longitude, latitude, and vessel speeds. As emphasized prior, the degree difference in the dataset frequency can, unfortunately, provide considerable uncertainty between the datasets once interpolation is applied. In this case, weather information is only recorded on an hourly basis; thus, the actual dynamic situation for each voyage condition may not be suitably represented. Additionally, on brief occasions, irregularities within the onboard sensor datasets are noticed. Seemingly, when the signals are lost, a corresponding null placeholder data entry is obtained instead of valid measurements. Unfortunately, such recording errors are tough to locate and can significantly skew the modelling analysis. Therefore, only by carefully applying the data preparation procedure can valid data points be ascertained from extreme interpolation and sensor errors.

Once the data has been aligned and cross-compared to validate the sets, the data cleaning can be conducted to eliminate irregular or outlier data points. However, before continuing, the combined dataset must be orientated towards each individual powering focus first. Since the proposed solution is to develop models considering both propulsion shaft power and auxiliary power, the associated data must reflect these different outcomes. Further outlined in section 4, a vessel operation classification is set to distinguish between Sailing and Anchor operations. Thus, the corresponding datasets can be likewise divided using this data feature. This clear division of functions allows for three datasets orientated towards each independent operation: Sailing, Anchor, and a Combined operation. A summary of each corresponding dataset operation and the associated data cleaning process effects can be seen in Table 1, where the amount dropped is highlighted red for each cleaning step.

In addition to the outlier detection approach, self-established cleaning specifications were incorporated to ensure physically feasible data solutions were retained. The corresponding data elimination criteria are identified as follows.

- Data <0 (excl. T_{air})
- $T_{air} > 40$ °C
- Radial°> 360°
- Rel. Humidity >95%
- Speed > $\text{Speed}_{\text{max}}$
- Power > $\text{Power}_{\text{max}}$

Ultimately, these specifications require much insight and engineering application knowledge. Based on the complete integration and cleaning process, it can be seen that while many data points are initially collected, a large portion of those is not useable. When considering the missing data, specifications, and outlier data, the

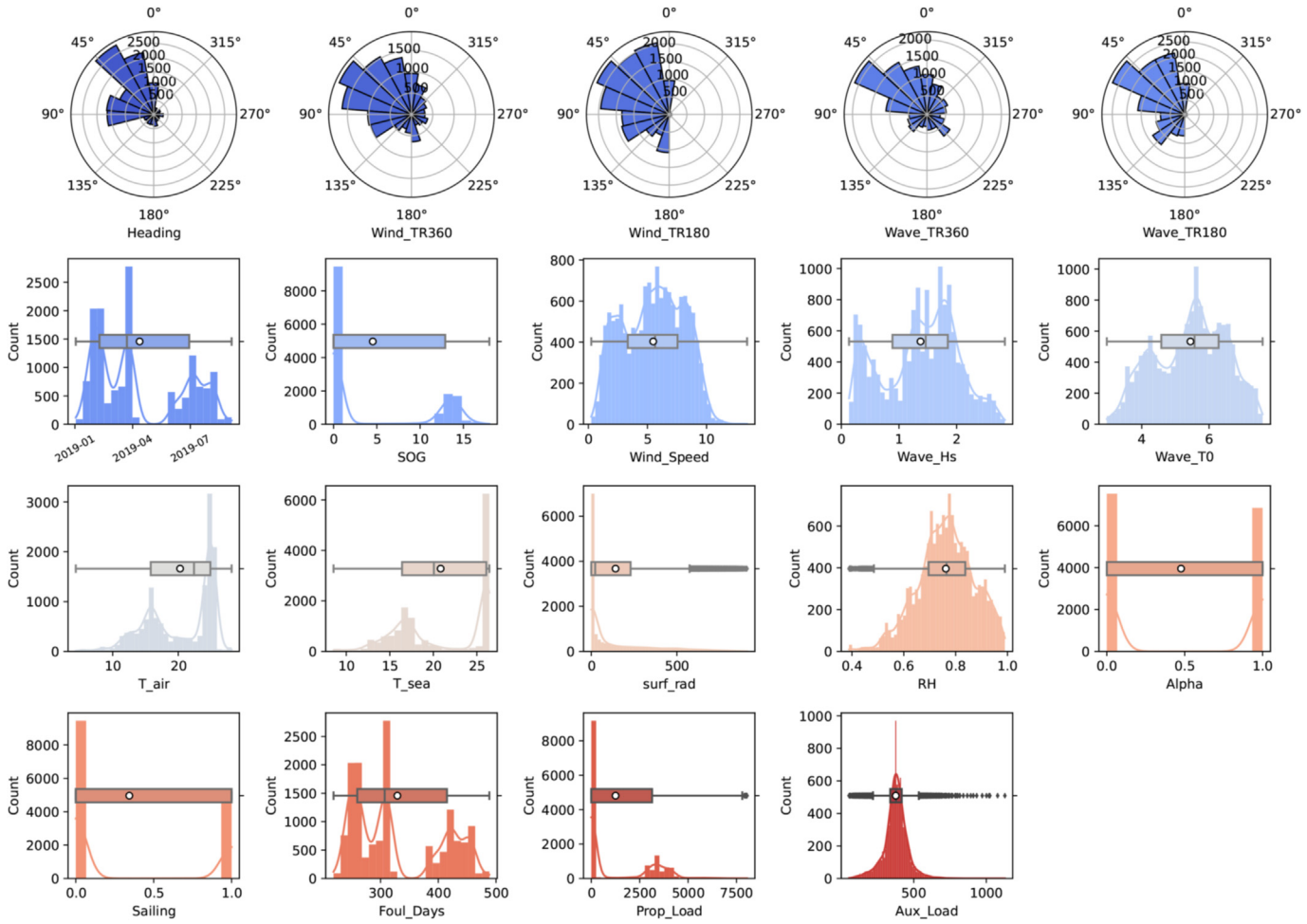

Fig. 4. Operational polar, histogram, and box-plot dataset input feature visualization.

Table 1
Dataset preparation summary.

Initial Datapoints	61944	
Pre - Processing Step	Amount Dropped	Data Remaining
Sailing Operation		
Sailing Only	55999	5945
Missing Data	993	4952
Data Specifications	111	4841
Outlier Drop	1245	3596
Final Datapoints	**58348**	**3596**
Anchor Operation		
Anchor Only	5945	55999
Missing Data	14506	41493
Data Specifications	32177	9316
Outlier Drop	2431	6885
Final Datapoints	**55059**	**6885**
Combined Operation		
Anchor + Sail	0	61944
Missing Data	15512	46432
Data Specifications	33976	12456
Outlier Drop	1964	10492
Final Datapoints	**51452**	**10492**

total data drop between the three operations was 39.5%, 87.7%, and 83.1%, respectively.

6.2. WBM evaluation and comparisons

Once the data has been cleaned, each WBM can be evaluated and directly compared with the recorded datasets. The total datasets for propulsion and auxiliary power demand and the associated relative proportions can be seen in Fig. 5.

When inspecting the propulsion comparisons, it can be seen that there is a relatively good correlation between the collected data results and the estimated predictions. The calm-water model generally presents a lower bound to the operational points. This relation is to be expected as this model only considers idealistic environmental conditions. When the additional resistance components are included, the total shaft WBM presents a much more realistic representation. However, there still exists a degree of difference between the gathered and estimation results. The model overestimates the dynamic impacts on power demand in the upper-speed regions, whereas the lower-speed areas underestimate the power demand.

Nevertheless, Fig. 5 indicates the mean proportion between the actual powering demand (blue) the estimated WBMs (orange) presents a marginal difference of 5%. The calm-water proportion (black) makes up the most considerable portion at 77%. The wave component likewise makes up a substantial proportion at 27% (beige). However, the overall influence of the wind model on propulsion power is at a mere 1% of the total recorded load.

The Auxiliary power comparison deviates quite drastically from

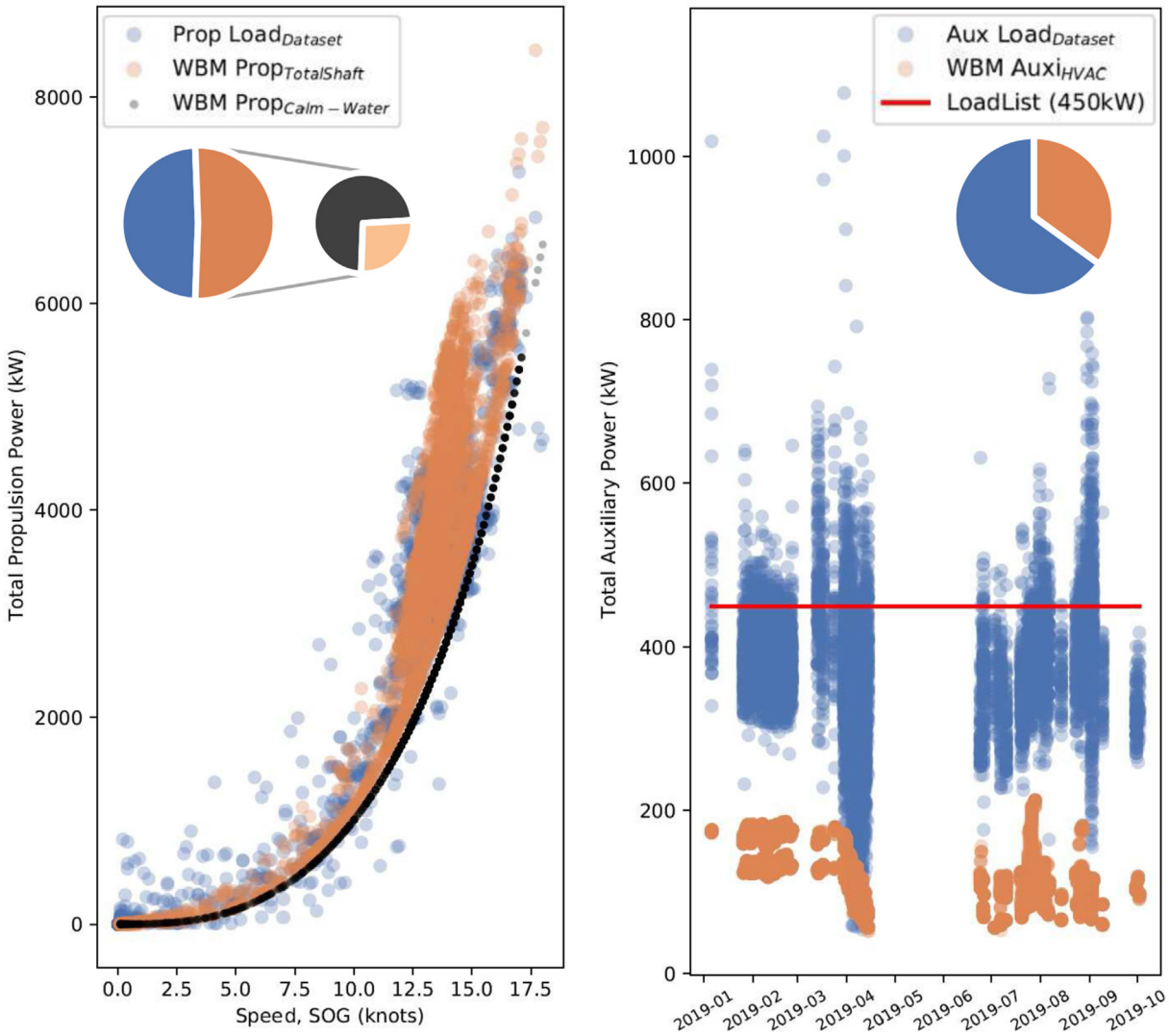

Fig. 5. Gathered operational target data and white-box model result comparisons between propulsion (left) and auxiliary power (right). Additionally, modelling decomposition of each mean modelling contributions is presented as proportions to the recorded target loads.

the propulsion WBM evaluations. It can be seen that the collected results have substantial variance (blue). As noted in section 3, the developed auxiliary WBM only considers the HVAC portion of the auxiliary power component, where an averaged proportion of 35% of the total load is observed (orange). Based on qualitative observation, the WBM does seemingly follow the general trending of the recorded auxiliary powers, thus capturing some dynamic effects. In addition to the developed auxiliary WBM, an existing operational load list (red) is used as a reference comparison benchmark. Unlike the WBM, this estimation represents the total expected load. Unfortunately, these calculations only provide a singular point-based solution, wherein the case vessel indicates a constant required load of 450 kW for successful operation.

It should be noted that on closer inspection of the WBM, a slight gap between estimations is present. This difference is due to the daylight factor, which influences the inherent radiation contribution. This factor quantifies the radiation component throughout a typical day within the localized timezone. The threshold considers the interior time between 06:00 to 18:00 as *Day* and the outer time as *Night*.

$$\alpha_{sun} = \begin{cases} 0 & t_{localhours} \leq 6 \text{ or } t_{localhours} > 18 \\ 1 & 6 < t_{localhours} \leq 18 \end{cases} \quad (5)$$

During the Night, radiation is not present; thus, the additional required power is neglected in such situations. Nevertheless, while not perfect and having no alternative other than a singular load list estimate, the WBM HVAC power model estimates a large dynamic portion of the total demand.

6.3. Feature Selection and optimization

Once all features are prepared and WBM(s) are evaluated, the Spearman correlation method is implemented as detailed in section 4. A visualization of the resulting combined operational rank matrix is presented in Fig. 6. Additionally, the results summary of the top 10 correlators and the mean absolute Spearman correlation for each dataset variation can be seen in Table 2.

When comparing the various operational conditions, it is clear that the auxiliary power dependencies are much lower than that of the other cases. Since these power systems are typically composed

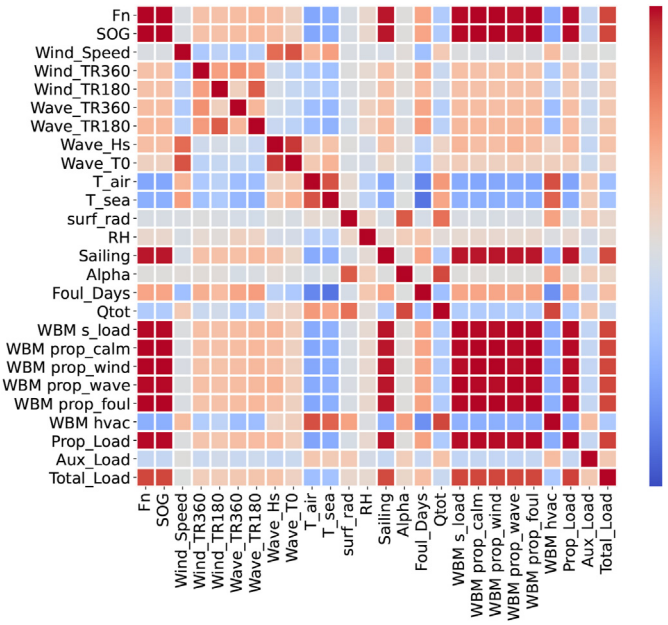

Fig. 6. Combined operation Spearman correlation rank matrix.

Table 2
Spearman correlation results summary for all operational conditions.

Sailing & Anchor:	Propulsion (IQR)		Auxiliary (IQR)	
Rank	Input	*SCo*	Input	*SCo*
1	V_s	+0.64	P_{hvac}[a]	+0.33
2	$P_{s,cw}$[a]	+0.63	E_e	+0.31
3	T_{air}	−0.63	α_{sun}	+0.24
4	$P_{s,t}$[a]	+0.53	T_{air}	+0.15
5	T_0	+0.49	β_{360}	−0.07
6	*HCI*	+0.49	α_{360}	−0.06
7	V_{wi}	+0.35	*RH*	+0.03
8	H_s	+0.33	T_0	−0.02
9	α_{360}	+0.31	H_s	+0.00
10	β_{360}	+0.02	V_{wi}	+0.00
Target	P_s	1.00	P_a	1.00
MASCo (T10)	0.44		0.12	
Combined:	**Auxiliary (IQR)**		**Total (non-IQR)**	
Rank	Input	SCo	Input	SCo
1	P_{hvac}[a]	+0.30	V_s	+0.86
2	T_{air}	+0.23	$P_{s,cw}$[a]	+0.86
3	E_e	+0.21	$P_{s,t}$[a]	+0.86
4	α_{sun}	+0.17	S_{oper}	+0.84
5	H_s	−0.17	T_{air}	−0.43
6	V_s	−0.17	P_{hvac}[a]	−0.35
7	S_{oper}	−0.16	H_s	+0.34
8	α_{360}	−0.14	*HCI*	+0.33
9	β_{360}	−0.14	α_{360}	+0.25
10	T_0	−0.13	β_{360}	+0.21
Target	P_a	1.00	P_{Total}	1.00
MASCo (T10)	0.18		0.53	

[a] White-Box Model (WBM) evaluated input feature.

of many different complex components, the Spearman Correlation approach cannot indicate a strong dependency between the independent and dependent features once combined. For instance, a dominating cubic relation between vessel speed and shaft propulsion power is expected due to the underlying physical relationships. Therefore, both ship speed as well as the corresponding WBM's show significant correlations in all cases. However, in the case of the auxiliary power, no dominating feature is known due to the complex cumulative nature of the total auxiliary power. Amongst the collected data features, the HVAC WBM and atmospheric conditions show the most dynamic dependencies; however, these only make up a portion (approximately 35%) of the total loading. As such, the input-output target dependencies may not be universally related to one or two features but instead composed of multiple low-order contributions from various elements.

Looking into the effects of the outlier detection methods also provides insight into the modelling performance. It is expected that the IQR method supports strong internal correlations as noise that may skew the modelling results is removed. While this seemingly influences the Sailing and Anchor operation by improving the overall correlations, the Combined Total operation indicates a dependency reduction. Since the IQR method is a parametric approach, it relies on traits that exhibit Gaussian-like distributions. However, as highlighted in Fig. 4, the highest correlators, such as vessel speed, do not present this form when merged, potentially providing distortion within the cleaning process.

Ultimately, while the correlations indicate some numerical connection, not all features with low dependencies should be disregarded. For instance, wind (beta) and wave (alpha) directions commonly indicate low dependencies; however, the correlation results are highly unreliable. These parameters are orientated using a polar coordinate axis, thus are continuous by nature. As such, the corresponding results may or may not be monotonic. Therefore, a minimum threshold is established to ensure some degree of correlation. This established threshold is set at ±0.10. Anything below the limiting bound is individually evaluated and judged for suitability based on the specific operation and the corresponding functions.

Based on the Spearman correlations and all available features (see Fig. 6), an average of 10 data features are retained across all operational conditions. It should be noted that the Combined datasets naturally have more input features as both auxiliary and propulsion orientated data features are to be kept.

At this stage, the black-box modelling phase begins with grid search hyperparameter optimization outlined in section 3.2. A summary of each operational dataset's optimal parameters and structure can be found in Table 3.

Based on the outcomes, many of the BBM parameters and structures converge towards a similar solution space. This conformity is understandable, as many of the datasets exhibit an overlap of data features. Nevertheless, it should be noted that the Combined operational datasets reached the upper bounding limit of the grid search evaluation. Therefore, while the overall structure is improved, there is still room for further enhancement. However, while keeping in mind general time limitations versus the overall expected modelling gains, further optimization of the network structure has been neglected. Ultimately, the grid refinement process can be conducted for an infinite number of hyperparameter selections and permutations to determine the optimal global structure; however, the respective computational demand increase must be judged carefully.

Table 3
Hyperparameter optimization results summary.

Type	Sailing Propulsion	Anchor Auxiliary	Combined Auxiliary	Combined Total
Cross-Validation	3x10	3x10	3x10	3x10
Dropout	0	0	0	0
Epochs	100	100	100	100
Batch Size	32	32	32	32
Activation	ReLu	ReLu	ReLu	ReLu
Optimizer	Adam	Adam	Adam	Adam
Layers	2	2	2	2
Neurons	38	38	40	40

6.4. Model evaluation

Having prepared the associated situational datasets and finding the optimal topography, each GBM can be developed and compared directly between the alternative modelling categories. The data was split with 70% used to train the network, 15% to test the final network, and 15% was used for early-stop validation, a standard split for these applications (da Silva et al. (da Silva et al., 2011)). Additionally, in this case, an ensemble of 30 models was considered to create the associated empirical bootstrapped aggregation confidence intervals. The associated performance metric results for each powering case are summarized in Tables 4–7, where the visualization results can be seen in Figs. 7–10, respectively. Furthermore, the associated input and output feature data ranges for each developed model are listed in appendix A.

Ultimately, the GBM solution has the best performance metrics amongst the three modelling categories whereas, the WBM's are typically least effective within the dataset training ranges. Based on the modelling requirements, all GBM solutions successfully achieve the established accuracy conditions. The observed ranking of performance was as follows,

$$GBM \geq BBM > WBM$$

A direct comparison between the GBM and WBM indicated that the former had an average of 15% improvement over the latter. Whereas the performance between the GBM and BBM was either slightly improved (1%) or equal. When investigating the suitability of the combined added resistance models, it could be seen from the results that the calm water model consistently had slightly better results. This result provides strong evidence to suggest that the added resistance models slightly reduced the internal correlations

Table 4
IQR Sailing Propulsion performance summary comparisons between GBM, BBM and WBM.

Sailing Propulsion Power					
Test #	539				
Metric	$GBM{:}P_{s,t}$	$GBM{:}P_{s,cw}$	BBM	$P_{s,t}$	$P_{s,cw}$
R^2	0.947	0.949	0.947	-1.219	-2.485
MAE	73.16	72.77	73.54	600.63	910.23
RMSE	121.15	120.02	122.41	237.96	996.83
Percent Error					
MAPE	2.23%	2.21%	2.23%	17.34%	25.94%
RMSPE	3.52%	3.49%	3.55%	22.98%	28.80%
Lo95%	258.74	253.86	260.77	-	-
Up95%	255.07	250.38	264.31	-	-
Cover%	96.66%	95.73%	95.92%	-	-

Table 5
IQR Anchor Auxiliary performance summary comparisons between GBM, BBM and WBM.

Anchor Auxiliary Power				
Test #	1033			
Metric	$GBM{:}P_{hvac}$	BBM	$Load\ List$	P_{hvac}
R^2	0.299	0.300	-2.78	-39.18
MAE	28.19	28.13	70.06	256.07
RMSE	35.09	35.07	79.56	259.36
Percent Errors				
MAPE	7.40%	7.40%	19.64%	67.00%
RMSPE	9.17%	9.16%	20.84%	67.94%
Lo95%	61.03	60.37	-	-
Up95%	69.71	69.09	-	-
Cover%	92.84%	92.84%	-	-

Table 6
IQR Combined Auxiliary performance summary comparisons between GBM, BBM and WBM.

Combined Auxiliary Power				
Test #	1574			
Metric	$GBM{:}P_{hvac}$	BBM	$Load\ List$	P_{hvac}
R^2	0.307	0.308	-1.736	-18.415
MAE	36.82	36.72	81.66	250.13
RMSE	47.96	47.92	96.38	256.74
Percent Errors				
MAPE	10.35%	10.34%	24.83%	66.65%
RMSPE	12.95%	12.94%	25.82%	68.79%
Lo95%	86.47	85.66	-	-
Up95%	98.04	98.38	-	-
Cover%	93.01%	93.39%	-	-

Table 7
non-IQR Combined Total performance summary comparisons between GBM, BBM and WBM.

Combined Total Power					
Test #	1814				
Metric	$GBM{:}P_{s,t+hvac}$	$GBM{:}P_{s,cw+hvac}$	BBM	$P_{s,t}+LL$	$P_{s,cw}+LL$
R^2	0.995	0.995	0.994	0.902	0.893
MAE	61.36	60.61	63.20	293.61	348.16
RMSE	120.34	119.79	126.97	541.90	564.94
Percent Errors					
MAPE	6.56%	6.32%	6.37%	19.05%	20.16%
RMSPE	7.17%	7.13%	7.56%	32.06%	33.48%
Lo95%	276.88	266.81	287.19	-	-
Up95%	252.04	245.18	259.00	-	-
Cover%	97.74%	97.85%	97.79%	-	-

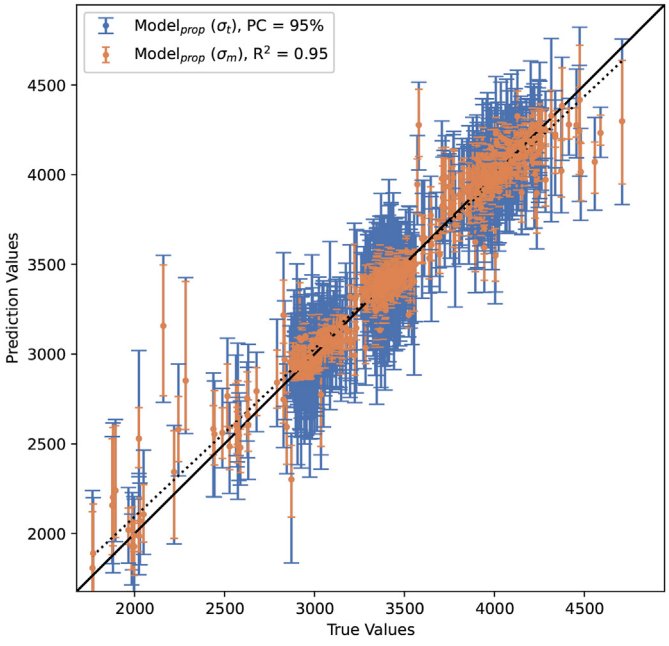

Fig. 7. IQR Sailing Propulsion $GBM + P_{s,cw}$ prediction performance with 95% confidence intervals.

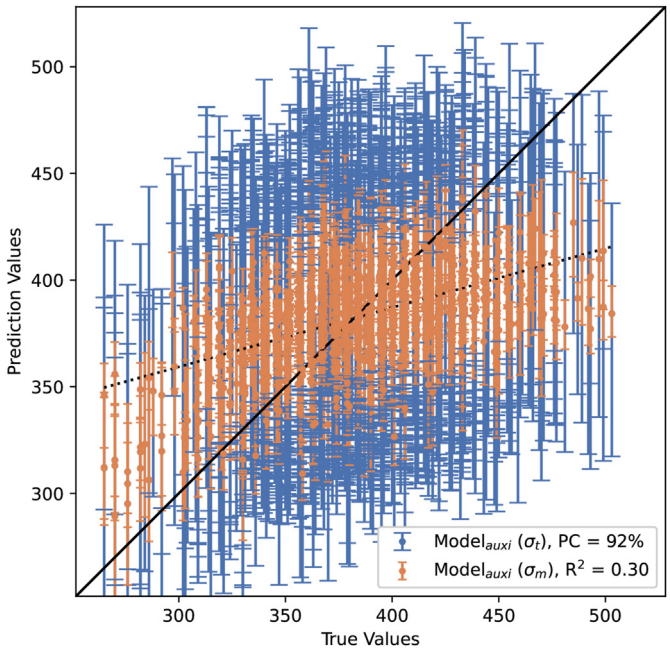

Fig. 8. IQR Anchor Auxiliary *BBM* prediction performance with 95% confidence intervals.

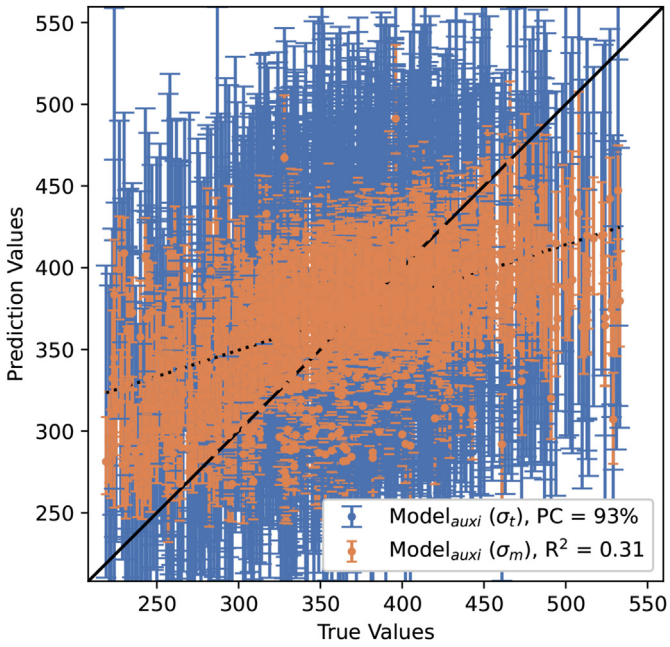

Fig. 9. IQR Combined Auxiliary *BBM* prediction performance with 95% confidence intervals.

meaning that the combined models did not improve in capturing the dynamic contributions of the measured total power. Nevertheless, the differences are marginal. However, a driving question remains why the global relative performance of the auxiliary power is consistently lower than the propulsion models. While there are no singular solutions, many aspects can contribute to the overall limited dynamic success.

First, there is a clear difference in Spearman Correlation magnitudes. It has been observed that the overall dependencies within the auxiliary cases are generally much lower than the operational

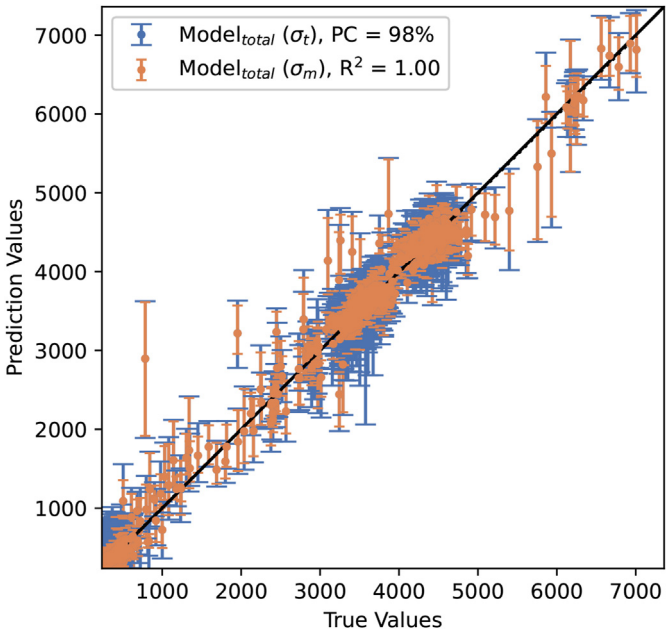

Fig. 10. None Combined Total *GBM* + $P_{s+cw\ +hvac}$ prediction performance with 95% confidence intervals.

propulsion equivalents. Therefore, as reflected within Table 2, the overall correlations prove to be significant initial gauges as to the expected performance. Lower dependencies mean that the BBM component of the GBM can not adequately map internal physical connections, thus resulting in a lowered goodness of fit.

Another reason, much related to the previous point, can be due to the HVAC WBM. It was hoped that this model would capture the most critical dynamic operational effects. While this model generally maintains the most significant correlations of all relevant data features, it is nowhere near the propulsion WBM contribution. This can mainly be attributed to the fact that overall auxiliary power is composed of multiple highly complex systems that continuously interact. Thus, the sizeable observed uncertainty can be related to the fact that, while the HVAC demand is generally considered the most prominent load proportion, the remaining system effects and influences are not considered. These critical missing components include; active stabilizers and rudders, which are known to contribute significantly to both auxiliary power demand and its dynamical operational contributions. For each corresponding best-developed model, the observed performance metric rankings were seen as follows,

Propulsion Sailing > Total Combined >
Auxiliary Anchor > Auxiliary Combined

While the auxiliary models do not behave as the propulsion models, the developed solutions are still below the established requirement thresholds. In a global sense, prediction accuracy within 35 kW of the total auxiliary load is still a significant improvement over the existing Load List estimation at a relative 80 kW difference.

While a model to evaluate total energy demand has been developed, this model cannot be easily decomposed into its various powering components. Unfortunately, this is a limitation of the GBM process, where a singular dependent target parameter is used to supervise the training process. However, proportioning can potentially be obtained using a traditional bottom-up approach. As such, a comparison between the aggregated propulsion and auxiliary models and the total consumption model is investigated.

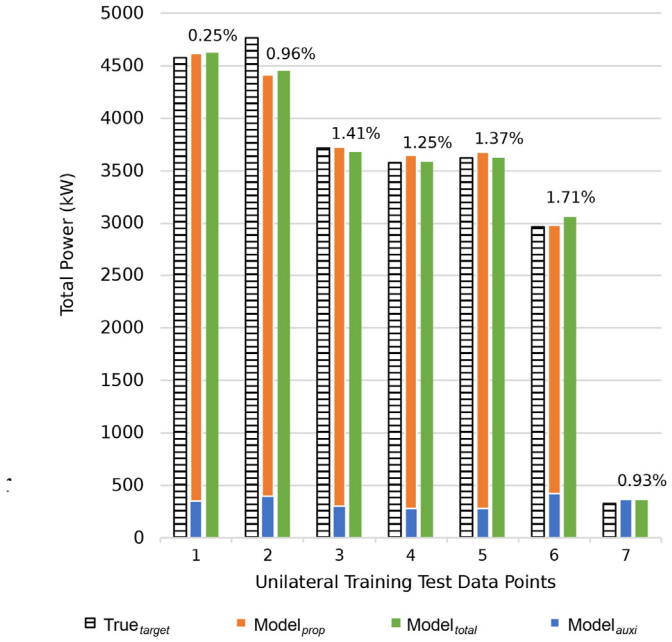

Fig. 11. Comparison between Total ($GBM + P_{s,cw\ +hvac}$) and Aggregation of Auxiliary ($\Psi\varnothing\ GBM + P_{hvac}$) + Propulsion ($GBM + P_{s,cw}$) power demand.

Fig. 11 highlights seven training test samples under varying conditions that have been used unilaterally between the three models. Here, the actual target, aggregated, and total power predictions are directly compared.

Based on the results, it can be confirmed that when both models are aggregated, the mean absolute difference between each sample case is marginal 1.1%. This deviation falls well below the mean error of each corresponding model and, as such, provides a suitable indication that a bottom-up GBM aggregation for system proportioning is a viable solution to capture each propulsion, auxiliary, and total powering demand. However, it should be noted that both the individual operational dataset range limitations and modelling uncertainties are retained and reflected in the combined model. Therefore, understanding each independent modelling capability with the aid of the developed confidence intervals is required to ensure the combined models behave appropriately.

7. Model verification

Having identified the best-developed models, a Verification analysis can be further conducted. This procedure ensures that the acquired output results are realistic, fall within similar order of magnitudes to other similar models, and align with the established modelling requirements. It should be enforced that the purpose is not direct accuracy comparisons as such models can vary depending on building factors, such as the amount of data. A detailed summary of the results and corresponding comparison results and outcomes are seen in Table 8.

It should be noted that while the absolute relative errors are transparent in each literature investigation, using only one performance metric, as previously indicated in section 4, does not adequately capture the total modelling outcomes. As such, the additional RMSE is implemented. However, since this metric is a dimensional unit, a normalization of the parameter must be conducted to compare the various modelling solutions unilaterally. As such, the following relation can be applied,

$$\text{nRMSE} = \frac{\text{RMSE}}{\bar{P}_{target}} \tag{6}$$

where the RMSE performance metric is normalized using the mean of the target parameters found within each literature investigation. While most of the literature sources are transparent in their results, not all reports have this metric listed. Thus, the average of the corresponding results is used as a comparison baseline for each operational condition. It can be seen that the related outputs from each case align within the same order of magnitudes of each literature result. The propulsion cases behave moderately better than most of the listed sources; however, this can also be attributed to the amount of data provided for model training. The auxiliary case also falls within the expected modelling outcome orders of magnitudes; thus, the overall modelling methodology can be verified.

7.1. Model extrapolation

Section 6.4 showed that the GBM and BBM models are highly effective within the developed model's training data ranges, where a high degree of prediction accuracy has been obtained. However, it can ultimately be noticed that both the GBM and BBM behave similarly with one another. Thus, a question remains as to why a more complex GBM modelling solution would be favoured over the more straightforward BBM approach.

The ultimate goal of the GBM is to introduce an aspect of the foundational physics attained from the WBM to aid in the internal dependencies and mappings. It is hoped that this integration allows for the prevention of unreasonable results when nearing and entering the extrapolation regions. Thus, to evaluate these

Table 8
Modelling results comparison summary.

Developed GBM Models	RMSE	$\bar{P}_{target}$	nRMSE	MAPE
Sailing Propulsion	121.145	3461.0	0.0350	2.2%
Anchor Auxiliary	35.094	381.8	0.0919	7.4%
Combined Auxiliary	47.959	373.2	0.1285	10.3%
Combined Total	120.343	1533.9	0.0785	6.3%
Propulsion				
Zwart(Zwart, 2020)			0.0842	6.63%
Parkes et al. (Parkes et al., 2018)			0.126	7.80%
Average			0.104	5.78%
Auxiliary (Buildings)				
Neto and Fiorelli (Neto and Fiorelli, 2008)			–	16.50%
Kalogirou and Bojic (Kalogirou and Bojic, 2000)			–	9.00%
Average			–	12.8%

unknown regions, data entries outside the training bounds were applied to the best-developed propulsion and auxiliary models. The additional data entries are collected from residual datasets, shown in Fig. 12, and the resulting performance summaries are presented in Tables 9 and 10, respectively. It should be noted that these results indicate a percent relative change from the performance of the interpolation and extrapolation models. In this instance, a simple classification between green (improved performance) and red (reduced performance) is used to identify gains and losses between associated performance metrics.

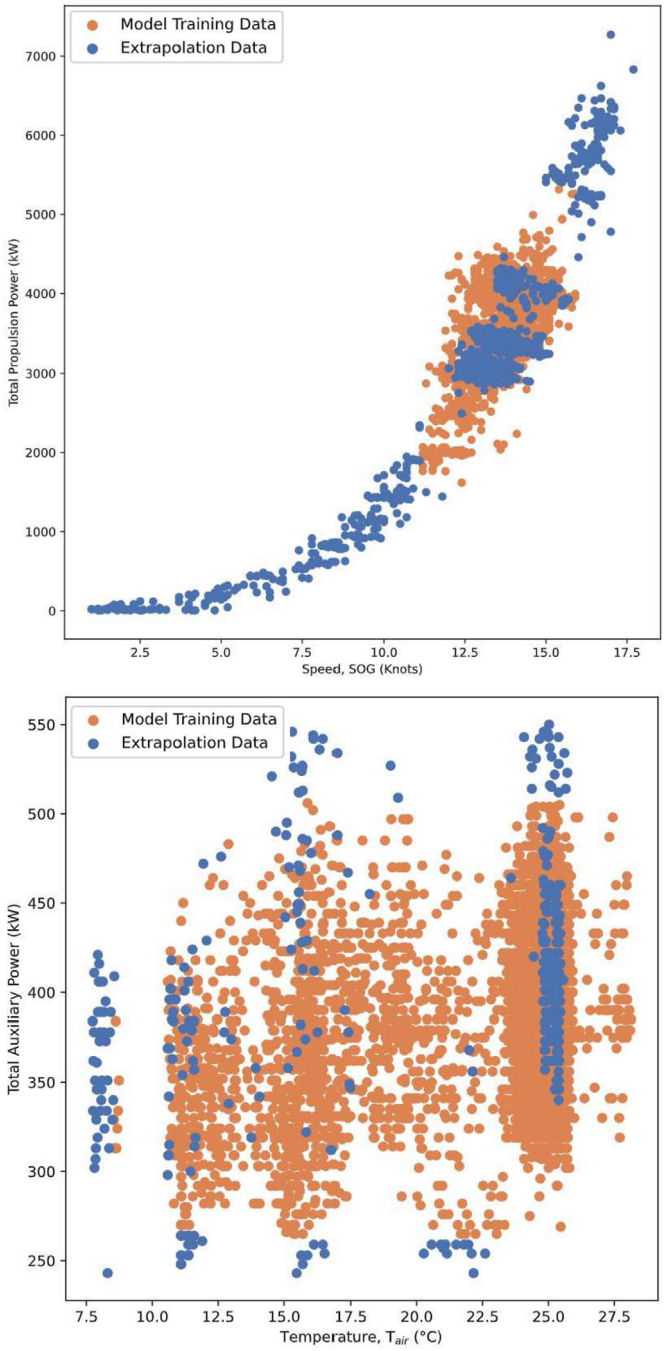

Fig. 12. Training versus extrapolation dataset comparison of propulsion (top) and auxiliary models (bottom).

Table 9
Sailing Propulsion extrapolation performance summary comparisons between GBM, BBM and WBM.

Sailing Propulsion Extrapolation					
Test #	1039				
Metric	*GBM:$P_{s,t}$*	*GBM:$P_{s,cw}$*	*BBM*	*$P_{s,t}$*	*$P_{s,cw}$*
R^2	0.722	0.788	0.694	0.83	0.704
MAE	416.86	373.15	445.59	491.40	728.57
RMSE	835.68	729.22	877.31	654.34	862.58
Propulsion Model Percent Relative Change					
ΔR^2	-24%	-17%	-27%	+168%	+128%
ΔMAE	+470%	+413%	+506%	-18%	-20%
ΔRMSE	+590%	+508%	+617%	+175%	-13%
Lo95%	1203.1	1146.23	1093.91	-	-
Up95%	2202.43	1991.77	2233.92	-	-
Cover%	93.74%	94.51%	93.36%	-	-

Table 10
Anchor Auxiliary extrapolation performance summary comparisons between GBM, BBM and WBM.

Anchor Auxiliary Extrapolation				
Test #	374			
Metric	*GBM:P_{hvac}*	*BBM*	*Load List*	*P_{hvac}*
R^2	0.285	0.284	-0.500	-13.58
MAE	47.627	48.55	72.369	269.31
RMSE	61.612	61.615	89.135	278.233
Auxiliary Model Percent Relative Change				
ΔR^2	-4%	-5%	+82%	+65%
ΔMAE	+69%	+73%	+3%	+5%
ΔRMSE	+76%	+76%	+12%	+7%
Lo95%	102.84	103.98	-	-
Up95%	127.44	121.41	-	-
Cover%	94.12%	94.39%	-	-

From a propulsion perspective, it can be seen that the WBM solution exhibits the best overall performance. This is to be expected, as one of the most significant benefits of using a WBM is its rooted fundamentals, allowing for improved extrapolation. However, it can be seen that the GBM is the next best performing model, whereas the BBM is the global worst performer by a substantial margin. As such, it can be clearly observed that the GBM propulsion model indicated a significant learning degree, demonstrating an improved extrapolation capacity compared to the BBM by approximately a 20% difference in MAE. Ultimately, this improvement can be related to the internal feature dependencies. The WBM data features all exhibited a high degree of correlation. Thus, solid internal mappings between the physics model and the BBM portion could be made.

Unlike the propulsion case, the auxiliary models behave radically differently. The GBM and BBM solutions, while a slight reduction in performance is seen, exhibit the best overall extrapolation performance. The WBM models, on the other hand, were inferior models to begin with, and as such, not much of a performance reduction be directly seen. In this case, the GBM shows a mere improvement of 2% MAE over the BBM. As such, this alludes to the unfortunate conclusion that the GBM was not successfully able to retain much of the physics induced by the integrated WBM. While there is a slight improvement over the BBM model, the overall performance is relatively weak to begin. Again, these

mediocre results can be related to each feature's internal correlations to the target power parameter. While the WBM did consistently have the most significant correlation, the magnitude of these relations is much lower and spread out than the propulsion cases. Thus, it can be suggested that limited relationships could be adequately developed to aid the model's ability to extrapolate beyond its design ranges.

8. General limitations and uncertainties

These general results provide critical insight into GBM interpolation and extrapolation behaviour. Ultimately, when correlations between the WBM data feature and the target parameter are significant, inherent physics retention can be achieved. This learning ability allows for a clear improvement of general performance and approaches the WBM's enhanced ability to extrapolate. However, when the parameter correlations are low, no such learning is achieved, and the modelling behaviour performs similarly to a pure BBM. While intuitively, this makes sense, it enforces the need for appropriate WBMs where the dynamic contributions instead of direct accuracy play the most critical role.

Unfortunately, this can link to a potential limitation in the proposed modelling approach. If data is continuously collected, there is a risk of developing a worse model if the correlation strength between the WBM and dataset decreases as trends over time change. Therefore, we should not only be concerned with increasing the data but having the most up-to-date models. Thus, the need for not just large data quantities but also suitable data features and up-to-date models that capture the target parameter's fundamental nature is essential.

Additional uncertainties within the modelling approach can also be related to potentially missing input parameters. Up until this point, the analysis has considered and assumed that all observed input and targeted points are independent and identically distributed (i.i.d). However, since the observations are recorded over time, the datasets have a clear temporal relationship. This temporal dependency, in effect, can influence the overall estimation for future model usage. Thus, an interesting consideration would be adding input features, such as Acceleration and Deceleration components, for instance. These use past data to quantify the corresponding present response. This input addition may help with the incorporation of some temporal information.

9. Conclusion

The following studied investigated a grey-box modelling approach using artificial neural networks and routinely applied white-box models to estimate dynamic operational powering demand. The results showed that the method significantly improved the estimation capacity of both propulsion and auxiliary power demand compared to a pure white-box model.

It can be concluded that the four method requirements have been confidently achieved. Section 6.4 demonstrated the incredibly accurate modelling evaluation capabilities for each corresponding operational consideration, all of which fell below the required 15% threshold. Furthermore, section 6.4 demonstrated the flexible nature of a GBM solution, allowing for simple model aggregation and proportioning based on a singular output using a multiple model scenario. In addition, and falling in line with the framework's flexibility, section 7.1 proved WBMs can be included to aid extrapolation by improving the input-output learning capacity if suitable models are selected. Finally, while not fundamentally present within the models, the established modelling methodology seen in section 4 considers and outlines proven techniques to handle irregular data entries and improve interpretability.

Ultimately, the developed models and the proposed modelling methodology has been inherently developed to handle and deal with growing datasets. As such, the approach should be able to quickly scale to other vessels within the *Feadship* fleet. While this broadened dataset may allow for improved application, it can also create sparse data regions. Additionally, new data features such as Froude number, ship length, displacement, or hull shape coefficients might be required to link the GBM inputs with the associated target outputs when including more vessels. These added features have the potential to increase complexity and thus, introduce the risk of overfitting. Nevertheless, supposing the domain is sufficiently broad and fully populated with the correct input data features, no evidence suggests that the approach would not be capable of meeting the required method constraints. Therefore, it is highly suggested to expand the current investigation to consider additional vessels within the *Feadship* fleet to maximize the application potential.

In conclusion, the grey-box modelling approach could improve the performance analysis of yachts and, therefore, can be used in all manners of design and operational stages, such as early-spiral iteration and real-time powering optimization, respectively. However, further improvements to the grey-box modelling approach are necessary to address limitations and uncertainties in the modelling approach. To drive these enhancements a series of future considerations and questions are presented as follows.

1. Feature selection sensitivity study: *Does the amount of selected top features influence the modelling and extrapolation performance?*
2. Variable WBM fidelity study: *Does using alternative WBM's improve modelling accuracy and extrapolation performance if the degree of WBM accuracy increases?*
3. Further operational profile decomposition: *Can more operational data splits enhance the performance and interpretability of the models?*
4. Multi-vessel incorporation study and comparison: *Can more vessels be added using the same workflow, and can similar performance be achieved?*
5. Temporal domain (time-series model) inclusion and comparison: *Does the selected modelling approach account for temporal influences, or does an alternative approach need to be considered entirely?*

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was performed as part of the MSc. Thesis for lead author, Odendaal (2021). The thesis was performed in Marine Technology at Delft University of Technology. The authors would like to acknowledge both the Delft University of Technology and De Voogt Naval Architects (Feadship) for their support of this research.

Appendix A. Developed Model Data Ranges

See tables 11, 12, 13, and 14.

Table 11
IQR Sailing Propulsion developed data range limits

Sailing Propulsion - IQR					
Details	Units	Minimum	Median	Mean	Maximum
V_s	knots	11.2	13.6	13.6	15.9
T_{air}	$^{\circ}C$	4.7	16.1	16.7	25.2
V_{wi}	knots	0.2	5.8	5.8	11.1
β_{360}	$^{\circ}$	0.0	117.7	153.0	359.8
H_s	m	0.2	1.9	1.8	2.8
T_0	s	3.3	6.0	5.9	7.6
α_{360}	$^{\circ}$	0.0	184.5	181.9	359.9
HCI	Days	286.3	311.8	340.6	459.3
$P_{s,cw}$	kW	1408.0	2543.5	2553.5	4213.5
$P_{s,t}$	kW	1581.5	3781.8	3804.2	6695.3
P_s	kW	1620.0	3406.0	3461.0	5316.0

Table 12
IQR Anchor Auxiliary developed data range limits

Anchor Auxiliary - IQR					
Details	Units	Minimum	Median	Mean	Maximum
T_{air}	$^{\circ}C$	8.6	24.7	22.8	28.1
RH	%	0.5	0.8	0.8	0.9
V_{wi}	knots	0.4	6.4	5.9	10.3
β_{360}	$^{\circ}$	0.0	62.5	102.6	360.0
T_0	s	3.0	5.6	5.5	7.5
H_s	m	0.2	1.4	1.3	2.3
α_{360}	$^{\circ}$	0.0	59.2	114.9	360.0
E_e	W/m^2	0.0	27.0	130.1	678.2
α_{sun}	Night/Day	0.0	0.0	0.4	1.0
P_{hvac}	kW	49.7	121.4	125.7	197.4
P_a	kW	265.0	379.0	381.8	506.0

Table 13
IQR Combined Auxiliary developed data range limits

Combined Auxiliary - IQR					
Details	Units	Minimum	Median	Mean	Maximum
V_s	knots	0.0	0.0	4.6	18.0
T_{air}	$^{\circ}C$	4.4	23.9	20.5	28.1
RH	%	0.5	0.8	0.8	0.9
V_{wi}	knots	0.2	5.9	5.7	13.4
β_{360}	$^{\circ}$	0.0	78.7	119.1	360.0
T_0	s	3.0	5.6	5.6	7.6
H_s	m	0.1	1.5	1.4	2.8
α_{360}	$^{\circ}$	0.0	77.5	135.5	360.0
E_e	W/m^2	0.0	11.7	97.5	573.2
α_{sun}	Night/Day	0.0	0.0	0.4	1.0
S	Anchor/Sail	0.0	0.0	0.4	1.0
P_{hvac}	kW	51.8	125.6	123.1	212.4
P_a	kW	219.0	375.0	373.2	535.0

Table 14
non-IQR Combined Total developed data range limits

Combined Total Power - None					
Details	Units	Minimum	Median	Mean	Maximum
V_s	knots	0.0	0.0	4.6	18.0
T_{air}	$^{\circ}C$	4.4	23.9	20.5	28.1
RH	%	0.5	0.8	0.8	0.9
V_{wi}	knots	0.2	5.9	5.7	13.4
β_{360}	$^{\circ}$	0.0	78.7	119.1	360.0
H_s	m	0.1	1.5	1.4	2.8
T_0	s	3.0	5.6	5.6	7.6
α_{360}	$^{\circ}$	0.0	77.5	135.5	360.0
HCI	Days	218.6	303.9	319.5	488.5
E_e	W/m^2	0.0	11.7	97.5	573.2
α_{sun}	Night/Day	0.0	0.0	0.4	1.0
S	Anchor/Sail	0.0	0.0	0.4	1.0
P_{hvac}	kW	51.8	125.6	123.1	212.4
$P_{s,cw}$	kW	0.0	0.0	859.8	6568.7
$P_{s,t}$	kW	0.0	0.0	1240.4	7699.3
P_{total}	kW	232.0	411.0	1533.9	7718.0

References

Aggarwal, C.C., 2018. Neural Networks and Deep Learning: A Textbook, 1 ed. Springer International Publishing AG.

Akershoek, R., 2018. Design of a Fully Electric Super Yacht: Research of Design to Find Driveres, Enablers, and Barriers of Electric Yachting. MSc. TU Delft Thesis.

ASHRAE, 2013. ASHRAE Handbook: Fundamentals. American Society of Heating. Refrigeration, and Air Conditioning Engineers.

Bal Beşikçi, E., Arslan, O., Turan, O., Ölçer, A.I., 2016. An artificial neural network based decision support system for energy efficient ship operations. Comput. Oper. Res. 66, 393–401. https://doi.org/10.1016/j.cor.2015.04.004.

Boertz, C., 2020. Energy Demand of a Fuel Cell-Driven Cruise Ship. MSc. TU Delft Thesis.

Botchkarev, A., 2018. Performance Metrics (Error Measures) in Machine Learning Regression, Forecasting and Prognostics: Properties and Typology, vol. 14. Interdisciplinary Journal of Information, Knowledge, and Management, 10.28945/4184. https://arxiv.org/abs/1809.03006.

Cepowski, T., 2020. The prediction of ship added resistance at the preliminary design stage by the use of an artificial neural network. Ocean Eng. 195, 106657. https://doi.org/10.1016/j.oceaneng.2019.106657, 10.1016/j.oceaneng.2019.106657.

Coraddu, A., Oneto, L., Baldi, F., Anguita, D., 2018. Vessels fuel consumption: a data analytics perspective to sustainability. Stud. Fuzziness Soft Comput. 358, 11–48. https://doi.org/10.1007/978-3-319-62359-7_2.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Boostrap. Springer-Science + Business Media B.V.

Ferrario, E., Pedroni, N., Zio, E., Lopez-Caballero, F., 2017. Bootstrapped Artificial Neural Networks for the seismic analysis of structural systems. Struct. Saf. 67, 70–84. https://doi.org/10.1016/j.strusafe.2017.03.003.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M., Herrera, F., 2016. Big data preprocessing: methods and prospects. Big Data Analytics 1, 1–22. https://doi.org/10.1186/s41044-016-0014-0.

Grin, R., 2014. On the prediction of wave-added resistance with empirical methods. Journal of Ship Production and Design 30, 1–11. https://doi.org/10.5957/JSPD.30.4.130060.

Holtrop, J., 1984. A statistical Re-analysis of resistance and propulsion data. Int. Shipbuild. Prog. 363, 272–276.

Holtrop, J., Mennen, G.G.J., 1982. An approximate power prediction method. Int. Shipbuild. Prog. 29.

ISO 7547, 2002. Ships and Marine Technology - Air-Conditioning and Ventiliation of Accomodation Spaces - Design Conditions and Basis of Calculations, vol. 2. ISO International Standard.

Kalogirou, S.A., Bojic, M., 2000. Artificial neural networks for the prediction of the energy consumption of a passive solar building. Energy 25, 479–491. https://doi.org/10.1016/S0360-5442(99)00086-9.

Kim, S.H., Roh, M.I., Oh, M.J., Park, W.W., Kim, I.I., 2020. Estimation of ship operational efficiency from AIS data using big data technology. Int. J. Nav. Archit. Ocean Eng. 12, 440–454. https://doi.org/10.1016/j.ijnaoe.2020.03.007.

Leifsson, L.T., Sævarsdóttir, H., Sigurdsson, S.T., Vésteinsson, A., 2008. Grey-box modeling of an ocean vessel for operational optimization. Simulat. Model. Pract. Theor. 16, 923–932. https://doi.org/10.1016/j.simpat.2008.03.006.

Man, Diesel, Turbo, 2006. Basic Principles of Ship Propulsion. Ship Propulsion.

Mazloumi, E., Rose, G., Currie, G., Moridpour, S., 2011. Prediction intervals to account for uncertainties in neural network predictions: methodology and application in bus travel time prediction. Eng. Appl. Artif. Intell. 24, 534–542. https://doi.org/10.1016/j.engappai.2010.11.004, 10.1016/j.engappai.2010.11.004.

Neto, A.H., Fiorelli, F.A.S., 2008. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. Energy Build. 40, 2169–2176. https://doi.org/10.1016/j.enbuild.2008.06.013.

Odendaal, K., 2021. Enhancing Early-Stage Energy Consumption Predictions Using Dynamic Operational Voyage Data: a Grey-Box Modelling Investigation. MSc. TU Delft Thesis.

Parkes, A.I., Sobey, A.J., Hudson, D.A., 2018. Physics-based shaft power prediction for large merchant ships using neural networks. Ocean Eng. 166, 92–104. https://doi.org/10.1016/j.oceaneng.2018.07.060.

Seakeeping Commitee of the 29th ITTC, 2018. Calculation of the weather factor fw for decrease of ship speed in wind and waves. International Towing Tank Conference.

da Silva, I.N., Spatti, D.H., Flauzino, R.A., Liboni, L.H.B.L., dos Reis Alves, S.F., 2011. Artificial Neural Networks. Springer Nature. https://doi.org/10.4324/9781315154282-3.

Stapersma, D., Klein Houd, H., 2012. Design of Propulsion and Electric Power Generation Systems. IMarEST.

Yang, L., Chen, G., Rytter, N.G.M., Zhao, J., Yang, D., 2019. A genetic algorithm-based grey-box model for ship fuel consumption prediction towards sustainable shipping. Annals of Operations Research URL. https://doi.org/10.1007/s10479-019-03183-5, 10.1007/s10479-019-03183-5.

Zwart, R.H., 2020. Trim Optimization for Ships in Service. MSc. TU Delft Thesis.