International Journal of Naval Architecture and Ocean Engineering 15 (2023) 100550

Contents lists available at ScienceDirect

International Journal of Naval Architecture and Ocean Engineering

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

Streamlined semi-automatic data processing framework for ship performance analysis

Prateek Gupta [a,*], Young-Rong Kim [a], Sverre Steen [a], Adil Rasheed [b]

[a] Department of Marine Technology, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[b] Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

ARTICLE INFO

Keywords:
Ship data processing
Ship performance analysis
Steady-state filter
Metocean hindcast interpolation
Ship hydrodynamics

ABSTRACT

The hydrodynamic performance of a sea-going ship can be analyzed using data from different sources, like onboard recorded in-service data, AIS data, and noon reports. Each of these sources is known to have its inherent problems. The current work highlights the most prominent issues, explained with examples from actual datasets. A streamlined semi-automatic approach to processing the data is finally outlined, which can be used to prepare a dataset for ship performance analysis. Typical data processing steps like interpolating metocean data, deriving additional features, estimating resistance components, data cleaning, and outlier detection are arranged in the best possible manner not only to streamline the data processing but also to obtain reliable results. A semi-automatic implementation of the data processing framework, with limited user intervention, is used to process the datasets here and present the example plots for various data processing steps, proving the effectiveness of the proposed approach.

1. Introduction

The performance of a sea-going ship is essential not only to keep the fuel and operational costs in check but also to reduce global emissions from the shipping industry. Analyzing the performance of a vessel is also of great interest for charter parties to estimate the potential of a vessel and the profit that can be made out of it. Therefore, driven by economic and social incentives, ship performance analysis and monitoring trade have been booming substantially in recent times. The importance of operational data from ships in this context is very well understood by most of the stakeholders, also reflected by the number of publicly funded research projects (like SFI Smart Maritime[1], which made this work possible) as well as the large number of industry partners involved in most of these projects.

The traditional way to evaluate the performance of a ship is using the noon report data provided by the ship's crew. A more exact approach, but not very feasible for commercial vessels, was suggested by Walker and Atkins (2007), conducting in-service sea trials in calm-water conditions regularly. With the advent of sensor-based continuous monitoring systems, the current trend is to directly or indirectly observe the evolution of the calm-water speed-power curve over time. ISO 19030 (2016), along with several researchers (Koboević et al., 2019; Coraddu et al., 2019) recommends observing the horizontal shift (along the speed axis) of the calm-water speed-power curve, termed as the speed-loss, over time to monitor the performance of a sea-going ship using the in-service data. Alternatively, it is suggested to observe the vertical shift of the calm-water speed-power curve, often termed as the change in power demand (adopted by Gupta et al., 2022; Carchen and Atlar, 2020). Some researchers also formulated and used some indirect performance indicators like fuel consumption (Koboević et al., 2019), resistance (or fouling) coefficient (Munk, 2016; Foteinos et al., 2017; Carchen and Atlar, 2020), (generalized) admiralty coefficient (Ejdfors, 2019; Gupta et al., 2021), wake fraction (Carchen and Atlar, 2020), fuel efficiency (Kim et al., 2021), power demand increase ratio (Guo et al., 2023; Mittendorf et al., 2023), etc. In each of these cases, it is seen (and most of the time acknowledged) that the results are pretty sensitive to the quality of the data used to estimate the ship's performance.

The ship's performance-related data usually inherits some irregularities due to several factors like sensor inaccuracies, the vibration of the sensor mountings, electrical noise, variation of environment, etc., as

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

* Corresponding author.
E-mail address: prateek.gupta@ntnu.no (P. Gupta).

[1] https://www.smartmaritime.no/.

https://doi.org/10.1016/j.ijnaoe.2023.100550
Received 13 January 2023; Received in revised form 7 September 2023; Accepted 7 September 2023
Available online 14 September 2023

pointed out in the Guide for Smart Functions for Marine Vessels and Offshore Units (Smart Guide) published by American Bureau of Shipping (2020). As presented by several researchers, it may be possible to carry out ship performance analysis using the data obtained from various sources, like onboard recorded in-service data (Gupta et al., 2022; Guo et al., 2023), publicly available AIS data (You et al., 2017; Kim et al., 2020a), and semi-autonomously recorded noon reports (Park et al., 2017). The quality of data used to carry out ship performance analysis and the results obtained further can be significantly improved by adopting some rational data filtering and processing techniques, as proposed by ISO 19030 (2016), Liu et al. (2020a), Kim et al. (2020b) and Guo et al. (2023). Therefore, while working towards evaluating the performance of a sea-going ship, it becomes quite clear that an easily adaptable data processing framework is required to process the data obtained from the ships-in-service. Addressing the same, Dalheim and Steen (2020a) presented a data preparation toolkit based on the in-service data recorded onboard two ships. The proposed toolkit was developed for a specific type of dataset, where the variables were recorded asynchronously and had to be synchronized before carrying-out ship performance analysis. The current work would instead focus on challenges faced while processing an already synchronized dataset as well as the data obtained from various sources, mentioned above.

The current paper introduces the reader to the problems generally faced while processing the datasets obtained from the ships during regular operation. As aforementioned, such datasets can be obtained from various sources, like onboard recorded in-service data, AIS data, and noon reports. Irrespective of the data source, the versatile data processing framework is developed to prepare the ship's performance-related datasets for ship performance analysis. Moreover, the data processing framework may be easily adapted for several other purposes, for instance, to help develop the concept of creating digital twins for sea-going ships, presented by Major et al. (2021). In such a case, the data processing framework can play an immensely critical role in preparing the data (if required, in real-time) for further use. Thus, the data processing framework presented here helps prepare datasets in a streamlined and semi-automatic manner for ship performance analysis or to act as a building block for technologies like digital twins for ships aimed at tasks like predictive maintenance, performance monitoring, ship autonomy, etc.

The following section discusses the art of ship performance analysis and the bare minimum characteristics of a dataset required to do such an analysis. Section 3 presents a quasi-steady-state filter found quite instrumental while processing the data. Section 4 presents the prominent challenges faced while processing the data from ships-in-service. Section 5 presents the data processing framework which can be used to process and prepare these datasets for ship performance analysis. Finally, section 6 finishes the paper with concluding remarks.

2. Ship's hydrodynamic performance analysis

The hydrodynamic performance of a ship-in-service can be assessed by observing its current propulsive state and, then, comparing it to a benchmarking standard. There are several ways to establish (or obtain) a benchmarking standard, like model test experiments, sea trials, CFD analysis, etc. It may even be possible to set a benchmarking standard using the in-service data recorded onboard a newly built ship, as suggested by Coraddu et al. (2019) and Gupta et al. (2021). Other than the benchmarking standard, the performance assessment methodology also requires recently recorded relevant data from the ship in operation, depicting the current propulsive state of the ship (which cannot be obtained using model test experiments or CFD analysis). Gathering such data requires near-real-time monitoring of the ship, and these datasets are known to have several sources of error, noise, and uncertainties, as discussed further in this paper. Therefore, evaluating the current performance of a ship requires a good amount of data processing. Moreover, the benchmarking standard is, generally, established for only a given environmental condition, most likely the calm-water condition. To draw a comparison between the current state and the benchmarking standard, the current performance must be translated to more or less the same environmental condition, therefore, increasing the complexity of the problem.

Assessing the hydrodynamic performance of a ship using the onboard recorded data is also challenging from the data collection, management, and transmission point of view. Firstly, collecting the data from ship-wide sensors into a centralized data acquisition system (DAQ) can be pretty problematic. The sensors may be adjusted to record measurements at different sampling frequencies, resulting in unsynchronized data values (as discussed by Dalheim and Steen, 2020a). Secondly, the recorded data may be too big to store as well as transmit to an onshore control and analysis center (as discussed by Perera and Mo, 2018). Therefore, decisions must be taken regarding sampling frequency, features (or variables) to be recorded, and adopted data processing methods, as discussed further in this paper.

2.1. Bare minimum variables

For translating the current propulsive state data to the benchmarking standard's environmental condition, and carrying out a reliable ship's hydrodynamic performance analysis, a list of bare minimum variables must be recorded (or observed) at a reasonable enough sampling rate. The bare minimum list of variables must provide the following information about each sampling instant for the ship: (a) Operational control, (b) Loading condition, (c) Operational environment, and (d) Operating point. The variables containing the above information must either be directly recorded (or observed) onboard the ship, collected from regulatory data sources such as AIS, or may be derived using additional data sources, like the operational environment can be easily derived using the ship's location and timestamp with the help of an appropriate weather hindcast (or Metocean) data repository.

The operational control information should contain the values of the propulsion-related control parameters set by the ship's captain on the bridge, like shaft rpm, rudder angle, propeller pitch, etc. The shaft rpm (or propeller pitch, in the case of ships equipped with controllable pitch propellers running at constant rpm) is the most crucial variable here as it directly correlates with the ship's speed-through-water. It should be noted that even in the case of constant power or speed mode, the shaft rpm (or propeller pitch) continues to be the primary control parameter as the set power or speed is actually achieved by using a real-time optimizer (incorporated in the governor) which optimizes the shaft rpm (or propeller pitch) to get to the set power or speed. Nevertheless, in case the shaft rpm (or propeller pitch) is not available, it may be appropriate to use the ship's speed-through-water as an operational control parameter, as done by several researchers (Liang et al., 2019; Farag and Ölçer, 2020; Minoura et al., 2021; Laurie et al., 2021; Mittendorf et al., 2023), but in this case, it should be kept in mind that, unlike the shaft rpm (or propeller pitch), the speed-through-water is a dependant variable strongly influenced by the loading condition and the operational environment.

The loading condition should contain information regarding the ship's fore and aft draft, which can be easily recorded onboard.

Table 1
The list of bare minimum data variables required for the ship's hydrodynamic performance analysis.

Category	Variables
Operational Control	Shaft rpm, Rudder angle, Propeller pitch
Loading Condition	Fore and aft draft
Operational Environment	Longitudinal and transverse wind speed, Significant wave height, Relative mean wave direction, Mean wave period
Operating Point	Shaft power, Speed-through-water

Although the wetted surface area and under-water hull form are more appropriate for a hydrodynamic analysis, these can be derived easily using the ship's hull form, if the fore and aft draft is known. The operational environment should at least contain variables indicating the intensity of wind and wave loads acting on the ship, like wind speed and direction, significant wave height, mean wave direction, mean wave period, etc. Finally, the operating point should contain information regarding the speed-power operating point for the sampling instant. Table 1 presents the bare minimum variables required for the ship's hydrodynamic performance analysis. The list given in the table may have to be modified according to ship specifications, for example, the propeller pitch is only relevant for a ship equipped with a controllable pitch propeller.

2.2. Sampling frequency

Almost all electronics-based sensors are known to have some noise in their measurements. The simplest way adopted to subdue this noise is by taking an average over several measurements (known as a 'sample' in statistics), recorded over a very short period. It is also believed that the statistical mean of a 'sample' converges to the true mean (i.e., the mean of the entire population), thereby subduing the noise, as the number of measurements in the 'sample' increases, provided the observations follow a symmetrical distribution. Averaging a certain number of samples has a similar effect on the high-frequency noise in the data as low-pass filtering. The more samples are averaged, the lower the effective cut-off frequency. The averaging technique, also known as down-sampling, has the benefit, compared to low-pass filtering, that it reduces the number of samples and, therefore, the amount of data, and it is computationally favorable. However, applying a low-pass filter before time-averaging the samples would probably help reduce noise in a much better way. Thus, time-averaging the data over short durations can be used to subdue noise, but it is still critical to decide on an appropriate or ideal sampling frequency.

The perfect sampling frequency would depend on the objective of the analysis and the recorded variables. For example, if the aim of the research is to predict the motion response of a ship or analyze its sea-keeping characteristics, the data should be recorded at a high enough sampling frequency such that it can capture such phenomenon. Hansen et al. (2011) analyzed the ship's rudder movement and the resulting resistance. They demonstrated that if the sampling interval were large, the overall dynamics of the rudder movement would not be captured, resulting in a difference in resistance. One criterion for selecting the data sampling rate is the Nyquist frequency (Jerri, 1977), which is widely used in signal processing. According to this criterion, the sampling frequency shall be more than twice the frequency of the observed phenomenon to capture the information regarding the phenomenon sufficiently. In some cases, the data logging system also applies a low-pass filter (for frequencies lower than half the Nyquist frequency) to avoid high-frequency noise before the sample value is recorded. However, if the aim is not to record any information regarding the above-mentioned moderately varying effects (instantaneous incident wind and waves, response motions, etc.), it may be acceptable to just obtain low-frequency time-averaged values so that such effects are subdued. Nevertheless, it may still be helpful to get high-frequency data or even just obtain the standard deviation or variance corresponding to each time-averaged sample. This can be advantageous from a data cleaning as well as condition monitoring point of view. For example, the legs of the time series showing very high variance, due to the noise or moderately varying effects, can be removed from the analysis to increase the reliability of results. Furthermore, the high-frequency data or the variance (corresponding to each time-averaged data sample) can be used to diagnose a technical failure or problem in equipment onboard the ship.

2.3. Sampling duration

It may be possible to evaluate the performance of a ship with just a few good data samples. Still, it may be considered unacceptable as the data samples may contain some noise, as discussed in the above section, resulting in a significant deviation from the actual value. Moreover, even if a singular data sample is obtained with very high confidence, one-point-performance evaluation may still be insufficient as it would provide the performance measure of the ship for a particular operating point (i.e., at only one value of speed and displacement), which may not be extendable to the whole operating range with a good enough confidence. Therefore, it is desired to evaluate the ship's performance by averaging over a large number of samples for several operating points, so that the averaging would make the results more resilient towards the noise, and the measured performance provides valuable information over a range of operating points with high confidence.

As aforementioned, the performance of a ship is evaluated by comparing the data from the ship-in-service with a benchmarking standard. The benchmarking standard is generally obtained for near-calm-water conditions in at least two loading conditions, i.e., laden (or design) and ballast. Thus, the collected data from the ship-in-service should also be obtained (or filtered) for near-calm-water conditions in similar loading conditions, to reduce the uncertainties introduced due to the corrections applied to the collected data for bringing it to the same conditions as the benchmarking standard. Also, evaluating the ship's performance for both the extremes of the loading condition, i.e., laden (or design) and ballast, may help interpolate the performance values for intermediate loading conditions with higher confidence. It should also be noted that some ships, like container liners, typically operate at more or less the same loading condition. In such cases, the performance evaluation is feasible as well as acceptable for just the same loading condition. Nevertheless, obtaining the data from a ship-in-service containing a handful of voyages may be enough to evaluate the performance of the ship, as long as the recorded data contains enough samples in near-calm-water conditions to have good coverage over the speed and displacement (or draft) range. However, the duration between these voyages should not be huge. Otherwise, the time-evolving nature of the ship's performance would influence the results.

2.4. Best practices

It is well-known that the accuracy of various measurements is not the same. It also depends on the source of the measurements. The measurements recorded using onboard sensors are generally more reliable as compared to the manually or semi-autonomously recorded noon report measurements, due to the possibility of human error as well as a much lower sampling frequency in the latter. Even in the case of onboard recorded sensor measurements, the accuracy varies from sensor to sensor and case to case. Some sensors can be inherently faulty, whereas others can give incorrect measurements due to unfavorable installation and operational conditions. Even the best ones are known to have some measurement noise. Thus, it is recommended to establish and follow some best practices for a reliable and robust ship performance analysis.

The onboard measurements for shaft rpm and shaft torque are generally obtained using a torsion meter installed on the propeller shaft, which is considered reliable, particularly for shaft rpm. The shaft power measurements are also derived from the same as the shaft power (P_s) is related to the shaft rpm (n) and torque (τ) through the following identity: $P_s = 2\pi n\tau$. It should be noted that no approximation is assumed in this formulation, and therefore, it should be validated with the data, if all three variables (n, τ, P_s) are available. The measurements for speed-through-water, on the other hand, are known to have several problems, as presented by Dalheim and Steen (2021). Thus, it is recommended to use shaft rpm (and not speed-through-water) as the independent variable while creating data-driven regression models to predict the shaft power. For the same reason, it may also be a good idea to quantify the

change in a ship's performance in terms of change in power demand rather than speed-loss (or speed-gain), which may be contrary to the speed-loss-based performance evaluation recommended in ISO 19030 (2016).

Further, it is also quite common to use fuel oil consumption as a key performance indicator for ship performance analysis (Karagiannidis and Themelis, 2021). The fuel oil consumption can be easily calculated from the engine-delivered torque and engine rpm if the specific fuel consumption (SFC or SFOC) curve for the engine is known. Even though the SFC curve is established and supplied by the engine manufacturer, it is only valid for a specific operating environment, and it is known to evolve over time due to engine degradation and maintenance. Thus, including fuel oil consumption in ship performance analysis increases the complexity of the problem, which requires taking engine health into account. If the objective of ship performance analysis is also to take into account the engine performance, then it may be beneficial to divide the problem into two parts: (a) Evaluate the change in power demand (for hydrodynamic performance analysis), and (b) Evaluate the change in engine SFC (for engine performance analysis). Now, the latter can be formulated as an independent problem with an entirely new set of variables of interest, like engine delivered torque, engine rpm, ambient air temperature, the calorific value of fuel, turbocharger health, etc. Alternatively, the change in SFC can be calculated using the engine fuel consumption data obtained by monitoring the onboard bunkering logs. Nevertheless, this two-part approach would not only improve the accuracy of ship performance analysis, but it would also allow the user to develop a more comprehensive and, probably, accurate analysis model. Furthermore, if the objective of the analysis is to evaluate the total energy efficiency of the ship, then the fuel consumption, as well as the performance of auxiliary systems (diesel generators, boilers, etc.), must also be taken into account, as they cannot be assumed constant due to their varying operational state and performance degradation.

3. Quasi-steady-state filter

A steady-state can be defined as a state in which the observed parameter remains unchanged, and a quasi-steady-state is a state in which the observed parameter changes so slowly that it can be assumed constant or unchanged. A quasi-steady-state detection algorithm or filter can be useful when processing time series data. As demonstrated further in the current work (sections 5.3, 5.7, and 5.10), such an algorithm can automatically identify all the instances of change in the state of a system by simply analyzing the data time series, which is not practical to do manually if the time series is very long. The difference in the state of the system may be due to a sudden or gradual adjustment of the control parameters (autonomously or manually) or a failure in the system. It may be crucial to identify and rectify (or remove) these state-change instances for further analysis.

In general, the steady-state of a system or a parameter can be identified by a simple gradient check. However, the time series data obtained from a realistic application domain would undoubtedly contain some noise, which would render the simple gradient-based steady-state filter ineffective. Therefore, a slope or higher order gradient-based quasi-steady-state filter is suggested here which can be tuned as per the task at hand or the amount of noise in the data. The quasi-steady-state filter presented here is an adaptation, with minor improvements, from the steady-state identification algorithm developed by Dalheim and Steen (2020b). It is implemented here in the following two stages: (a) The first stage, which uses a sliding window to remove unsteady samples by performing a *t*-test on the slope of the data values, as suggested by Dalheim and Steen (2020b), and (b) An optional second stage, which performs an additional back-gradient check for the samples failing the *t*-test in the first stage. The first stage sliding window size as well as the critical limits for the t-tests (in the first and second stages) are user-defined and need to be tuned based on the data time series.

The suggested improvement to the original algorithm is that, unlike Dalheim and Steen (2020b), the t-value ($\mathbf{t}$) for the first stage *t*-test should be calculated as follows:

$$\mathbf{t} = \frac{\widehat{b}_1}{1 + \widehat{\sigma}_1} \tag{1}$$

where $\widehat{b}_1$ is the estimated slope of the sliding window, and $\widehat{\sigma}_1$ is the estimated standard deviation of the slope. Here, 1 is added in the denominator to avoid infinity when the standard deviation of the slope ($\widehat{\sigma}_1$) goes to zero. Thus, the improved first stage, now, would not misclassify the data points falling on a horizontal and almost straight line, with very small estimated standard deviation (i.e., $\widehat{\sigma}_1 \to 0 \implies \mathbf{t} \to \infty$), as unsteady. Here, the slope and its standard deviation can still be estimated using ordinary least squares (OLS) linear regression, as demonstrated by Dalheim and Steen (2020b).

In the second stage, the backward gradient ($\partial x_i/\partial t$) for all the samples failing the first stage *t*-test is calculated as follows:

$$\frac{\partial x_i}{\partial t} = \frac{x_i - x_{i-1}}{t_i - t_{i-1}} \tag{2}$$

where x_i is the value of the ith sample (which failed the first stage *t*-test), and x_{i-1} is the value of the sample just before x_i, regardless of whether x_{i-1} failed or passed the first stage *t*-test. t_i and t_{i-1} represent the time of observation for the ith and $(i-1)^{th}$ data sample. Finally, the absolute value of the backward gradient ($|\partial x_i/\partial t|$) is compared with a threshold value[2], and all the samples below the threshold value are added back to the quasi-steady samples' list. The second stage helps retain some samples, which would fail the first stage *t*-test as these samples lie at the starting or end of an unsteady leg, resulting in a high estimated slope. Nevertheless, it should be noted that the second stage is optional and can be skipped. If the data is very noisy, it is most definitely recommended to skip the second stage as it would not be appropriate to rely on simple gradients in this case.

Another problem, which may reduce the filter's effectiveness, can be due to highly non-uniform sampling intervals or missing data samples. In this case, the first stage sliding window width, which is defined by a fixed number of samples, may become quite large for some sections of the time series (where many samples are missing). This may result in misclassifying some of the unsteady sections of the time series as steady. In such a case, defining the sliding window width in terms of fixed time interval instead of the number of samples (suggested by Dalheim and Steen, 2020b) is found to produce better results. Here, the number of samples in the sliding window would vary as it slides forward due to the non-uniform sampling interval. Moreover, the degrees of freedom for the Student's t-distribution (used for the *t*-test) can be defined as the maximum number of samples that can be accommodated in the window.

4. Problems associated with ships' operational data

As mentioned earlier, the data required to evaluate the performance of an in-service sea-going ship can be obtained from various sources. The three primary sources are: (a) onboard recorded in-service data; (b) AIS data; and (c) noon reports. Each of these data sources has its inherent problems and some problems which can be found in all the data sources. The most prominent issues are discussed here.

4.1. Missing or insufficient information

In order to carry out an analysis, the available dataset must contain a bare minimum list of variables, containing ample information which may be required to model or understand the state of the phenomenon at

[2] The threshold for the backward gradient test can also be obtained using a user-defined significance level (α) for the Student's t-distribution, technically making it a *t*-test again.

a given point in time. The same is applicable to ship performance analysis. Section 2.1 presents the bare minimum variables required to model the hydrodynamic state of a ship. The first biggest challenge here is, therefore, to obtain the variables listed in Table 1 for each data sample. The onboard recorded in-service data generally contains most of these variables, with the exception of wave information, i.e., significant wave height, mean wave period, and relative mean wave direction. Some modern ships which are fitted with wave radars can even record wave information in real-time. Nevertheless, the weather information can be easily obtained from one of the publicly available weather hindcast (Metocean) data repositories, but interpolating the weather data variables from such repositories to the ship's location at a given time can be challenging. In the case of AIS data and noon reports, the list of available variables is much shorter, which presents an even bigger challenge.

Other than the data variables, a substantial amount of information may also be required regarding the object which is being observed. For instance, to carry out the performance analysis of a ship, information regarding the ship itself, like its principle particulars (or dimensions), hull form, design, etc., is most likely required. Such information is needed to derive or estimate additional variables necessary for further analysis. In the case of ship performance analysis, the information regarding the ship may be necessary to estimate the hydrostatic, hydrodynamic, and environmental loads acting on the ship, which are further used to estimate the total resistance acting on the ship. If the ship's information is unavailable, it may be possible to obtain it from a sister ship, standardized ship designs, or regression formulas based on standardized ship designs.

4.2. Faulty sensor installations

Some of the sensors installed onboard a ship can provide incorrect measurements due to improper installation. For instance, Wahl (2019) presented the case of faulty installation of the wind anemometer onboard a ship, resulting in missing measurements for head-wind conditions, probably due to the presence of an obstacle right in front of the sensor. Such a fault is reasonably simple to deal with, say, by fixing the installation of the sensor. It may even be possible to improve the already recorded data using the wind measurements from one of the publicly available weather hindcast data repositories. However, it is crucial to identify such problems using data exploration and validation techniques before carrying out any further analysis.

4.3. Measurement errors

The error in a sensor measurement can be seen as having two main sources: (a) noise; and (b) bias. The white noise observed in electronics-based sensors is discussed in detail in section 2.2. As mentioned in section 2.2, the sensor measurements are generally averaged over a short period to subdue the white noise. However, there may still be some remaining white noise in the recorded data. Thus, the analysis scheme should be designed such that the results are robust towards the presence of any white noise. The other sources of measurement error, i.e., systematic but irregular noise and biases, are not that easy to handle. The biggest challenge here is identifying such errors, as they can be quite unpredictable. It is, therefore, recommended to carry out a thorough examination of the dataset by employing as many data validation schemes as possible and visual or smart statistical data exploration. This would need user intervention which is prone to human error, but it may still produce some fruitful results. Moreover, a good knowledge of the application domain would also help while examining the dataset and finding anomalies.

It may also be possible to identify some of the systematic errors and biases by studying and understanding the shortcomings of the measurement sensors. The most commonly known defects are observed in the draft and speed-through-water measurement sensors. As pointed out by Gupta et al. (2021), pressure-based draft sensors are susceptible to systematic errors due to the so-called Venturi effect. The pressure transducer, in practice, measures the total pressure acting on the bottom plate of the ship at the location of the transducer, which is further converted into the corresponding water level height or the draft measurement. When the ship starts to move, the non-zero relative water velocity between the ship's bottom and water causes negative hydrodynamic pressure at the locations of the draft pressure sensors, and therefore, further measurements taken by the draft sensors are incorrect. This is known as the Venturi effect. Unfortunately, there is no established method to fix the draft measurements in such a case, as the localized hydrodynamic pressure at the transducer is difficult to estimate because it depends on both the speed-through-water and the local hull geometry around the pressure sensors. In addition to the Venturi effect, the relative water velocity (and reduced total pressure at the ship's bottom) also influences the actual draft of the ship, typically leading to a slight increase in the draft and bow-down trim. This effect, popularly known as the squat effect, becomes quite prominent in shallow water conditions due to the presence of the seabed. However, the squat effect should not be mistaken for the Venturi effect while correcting the in-service draft measurements. It should be noted that the former influences the actual draft and trim of the ship, while the latter only influences the draft measurements, i.e., it is a measurement error, which should be fixed before any further analysis.

The systematic errors and biases present in the case of the speed-through-water sensor are much more complicated. The state-of-the-art speed-through-water measurement device uses the Doppler acoustic speed log principle. Here, the relative speed of water around the hull (i.e., the speed-through-water) is measured by observing the frequency shift (popularly known as the Doppler shift) of the ultrasound pulses emitted from the ship's hull, due to its motion. The ultrasonic pulses are reflected by the ocean bottom, impurities in the surrounding water, marine life, and even the liquid-liquid interface between the density difference layers in the deep ocean. The speed of water surrounding the ship is influenced by the boundary layer around the hull so it is required that the ultrasonic pulses reflected only by the particles outside the boundary layer are used to estimate the speed-through-water. Therefore, a minimum pulse traveling distance has to be prescribed for the sensor. If the prescribed distance is too larger or if the ship is sailing in shallow waters, the Doppler shift is calculated using the reflection from the ocean bottom, i.e., the sensor is in ground-tracking mode, and therefore, it would clearly record the ship's speed-over-ground instead of the speed-through-water. On the other hand, if the minimum pulse traveling distance is set too short, the measurements can be affected by the boundary layer. Dalheim and Steen (2021) presented a detailed account regarding the uncertainty in the speed-through-water measurements for a ship, commenting that the speed log sensors are considered one of the most inaccurate ones onboard the ship.

4.4. Data outliers

Another big challenge with measurement data is the problem of detecting and handling outliers. An outlier is an anomalous data sample that does not follow the usual trend, observed in the remaining data samples. Although it may be possible to categorize outliers as measurement errors, the difference between an outlier and a measurement error, here, is that the former is assumed to be unsystematic and occurs due to an unexpected failure. Moreover, the failure resulting in outliers may be temporary and short-lived, or it may be a permanent sensor breakdown, which would need sensor adjustment, repair, or replacement. Gupta et al. (2021) observed that the recorded ship heading was filled with zeros in the latter part of the onboard recorded in-service data time series. This is probably due to a permanent sensor failure. Such a problem can be easily identified by carrying out proper data validation with visual or smart statistical data exploration. The problem of finding only a handful, but highly influential, outliers comfortably hidden in a

long data time series can be many-fold challenging.

As suggested by Olofsson (2020), it may be possible to categorize outlier samples into the following two broad categories: (a) Contextual outliers and (b) Correlation-defying outliers[3]. Dalheim and Steen (2020a) presented methods to detect and remove contextual outliers, further categorized as: (i) obvious (or invalid) outliers; (ii) repeated values; (iii) drop-outs; and (iv) spikes. Contextual outliers are easily identifiable as they either violate the known validity limits of one or more recorded variables (as seen in the case of obvious outliers and spikes) or present an easily identifiable but anomalous pattern (as seen in the case of repeated values and drop-outs). The case of correlation-defying outliers is much more difficult to handle, as they can easily blend into the cleaned data pool. The two most popular methods which can be used to identify correlation-defying outliers are Principal Component Analysis (PCA) and autoencoders. Both these methods try to reconstruct the data samples after learning the correlation between the variables. A correlation-defying outlier would result in an abnormally high reconstruction error and, therefore, can be detected using such techniques. In a recent attempt, Thomas and Judith (2021) demonstrated an ensemble method combining PCA and autoencoders coupled with isolation forests to detect such outliers.

4.5. Angular measurement error due to time averaging

The onboard recorded in-service data can be supplied as time-averaged values over a short period (generally up to around 15 min). Although the time-averaging method subdues noise in the data samples (as discussed in section 2.2), it introduces a new problem in the case of angular measurements. The angular measurements are, generally, recorded in the range of 0–360°. When the measurement is around 0 or 360°, it is evident that the instantaneous measurements, reported by the sensor, will fluctuate in the vicinity of 0 and 360°. Now, assuming that the sensor reports a value of about 0° for half of the averaging time and about 360° for the remaining time, the time-averaged value recorded by the data acquisition (DAQ) system will be around 180°, which is significantly incorrect. Most of the angular measurements recorded onboard a ship, like relative wind direction, ship heading, etc., are known to inherit this problem. It should be noted that, unlike the example given here, the incorrect time-averaged angle can take any value between 0 and 360°, depending on the instantaneous values over which the average is calculated. Although it may be possible to fix these incorrect values using a carefully designed algorithm, there is no established method currently available.

4.6. Uncertainty due to long-time averaging & human error

Generally, the information supplied through noon reports is obtained based on onboard sensor measurements and manually logged values. Here, the data collection interval is once a day, and most of the information is manually logged in as the average of the values observed or accumulated during the last 24 h, for instance, the distance traveled, the average speed of the ship, and fuel consumed in the last 24 h. Therefore, apart from the above-cited problems like sensor measurement errors, the noon report data may have problems due to the use of (24-h) long-time averaging and human error. Aldous et al. (2015) performed a sensitivity analysis to assess the uncertainty in ship performance analysis due to the uncertainty in the input information, using the data supplied as continuously recorded in-service data as well as the noon reports. It was observed here that the uncertainty in the results was significantly sensitive to the number of samples in the dataset. In other words, such uncertainty can be mitigated through the use of data representing longer time series, data collected with higher frequency, and data that is processed rationally. These results were also confirmed by Park et al. (2017) and Themelis et al. (2018). Park et al. (2017) demonstrated in a case study that the reported power or energy consumption between the noon reports and onboard recorded in-service data differed by 6.2% and 17.8% in ballast and laden voyage conditions, respectively.

Using the averaged values over a long period, as in the case of noon reports, the variations due to acceleration/deceleration and maneuvering cannot be captured (also discussed in section 2.2). Moreover, in the case of ships that sail relatively short voyages such as feeder ships and ferries, inappropriate noon report data may be obtained for performance analysis due to frequent changes in the operational state. Besides, regarding the weather and sea state information, the supplied information generally corresponds to the condition right before the noon report is sent from the ship. Therefore, it is not possible to account for the changes in the performance of the ship due to the variation in weather conditions during the last 24 h. Moreover, some of the information logged in the noon report is read and noted by a person from onboard sensor measurements. Here, it is possible that the time at which the values are read from the sensors every day may be different as well as different sensors may be used for the values to be logged-in for the same variable. Also, there may be cases when the observed value is incorrectly logged into the noon report, sometimes even with an intent to tamper with the data. Thus, if the process of preparing the noon reports is not automated, there will always be the possibility of human error and data tampering. Automated data recording systems, like onboard recorded in-service data and AIS data, can be considered more reliable, but they can also be tampered with in some unfortunate cases.

5. Results: data processing framework

The results here are presented in the form of the developed data processing framework, which can be used to process raw data obtained from one of the previously mentioned data sources, i.e., high-frequency in-service data, AIS data, and noon reports, for ship performance analysis. The data processing framework is designed to resolve most of the problems discussed in the above section. Fig. 1 shows the flow diagram for the data processing framework. The following sub-sections briefly explain the consecutive processing steps of the given flow diagram, and the last sub-section (5.11) presents results from an in-service dataset, processed using the proposed framework. It may be possible that the user may not be able to carry out some of these steps due to the unavailability of some information or features in the dataset. For example, due to the unavailability of the GPS data (latitude, longitude, and timestamp variables), it would not be possible to interpolate weather hindcast data. On the other hand, it may be possible that some of the steps suggested here may not be relevant in some cases. For instance, in order to use a completely data-driven approach, like machine learning (ML), calculating hydrostatics as well as resistance components may not be relevant, as the ML approach may not need these features (or variables) for creating the model. In such cases, it is recommended to skip the corresponding step and continue with the next one.

Semi-automatic Processing. The data processing framework has been outlined so that, after being implemented, it can be executed in a semi-automatic manner, i.e., requiring limited intervention from the user. The semi-autonomous nature of the framework would also result in fast data processing, which can be important for extensive datasets. The implementation of the framework in terms of executable code is also quite essential to obtain a semi-automatic and fast implementation of the data processing framework. Therefore, it is recommended to adopt best practices and optimized algorithms for each processing step according to the programming language being used. The data processing framework would also need many details regarding the ship and dataset. Creating standard templates for such information and importing them (as resources or libraries) into the main executable code while processing the data would help increase the level of autonomy. For instance, information like the ship's principle particulars, hydrostatics table, model test, sea trial data, and parameters for wind and wave resistance

[3] Called collective outliers by Olofsson (2020).

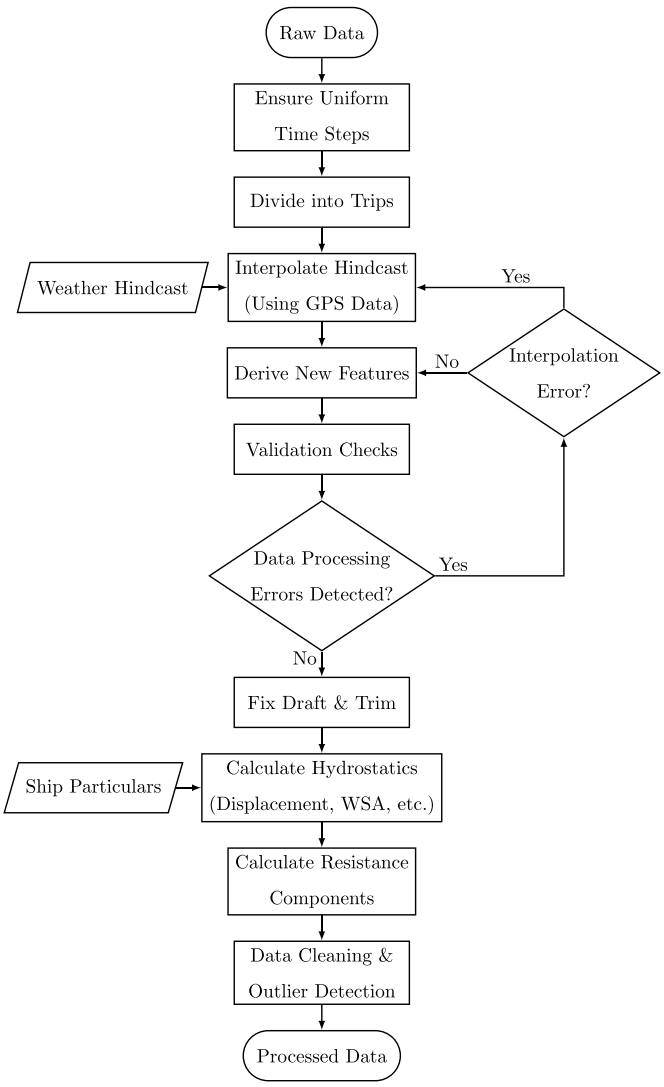


Fig. 1. Data processing framework flow diagram.

estimation, etc. can be standardized into CSV (Comma Separated Values) or equivalent formats for auto-importing and processing. If information regarding some parameters is not available, empirical methods to estimate these parameters can be employed, keeping in mind that adopting such strategies may increase the uncertainty in results. In the case of hindcast (or Metocean) data, the relevant data files can be downloaded automatically (or manually) on a local drive using a code/script exploiting an API (Application Programming Interface), made available by most of the hindcast data providers. In such a manner, the data processing can not only be automized to a better extent, but it would also enable the user to switch to the next ship smoothly.

Validation. The reliability of the data processing activity is also quite crucial to obtain good results. Therefore, it is essential to carry out the validation of work done in each processing step. The usual practice adopted here, while processing the data using the framework, is to create several visualizations, like time series plots of data variables in the trip- or voyage-wise manner (explained later in section 5.2) at the end of each processing step, and then, inspecting them to validate the outcome. It may also be possible to use some smart statistical data exploration and, or validation techniques instead of manually inspecting the visual plots as one might believe that manual inspection is prone to human error, but unfortunately, no such well-established method is known today. Moreover, the framework's effectiveness (as a whole) can also be tested by validating the performance predicted based on the data, processed using the given framework. Such a validation, although considered quite difficult, can be done by comparing the predicted performance with the measured hull roughness (when the ship's hull is inspected, say, during dry-docking) or the results obtained from the in-service sea trials, as suggested by Walker and Atkins (2007). However, it should be kept in mind here that the effectiveness of the framework also depends on the methodology adopted in each data processing step, for instance, the empirical or physics-based methods adopted for the estimation of added resistance components. Thus, decisions regarding the adoption of practices contributing to the data processing framework should be made after a thorough validation of these methods for the given ship, as discussed further in this paper.

5.1. Ensure uniform time steps

Ensuring uniform and evenly-spaced samples would not only make it easier to apply time-gradient-based data processing or analysis steps. It would also help avoid any misunderstanding while visualizing the data, by clearly showing a gap in the time series plots (even when the data is plotted against sample numbers) and removing any abrupt jumps in the data values. Depending on the data acquisition (DAQ) system, the in-service data recorded onboard a ship is generally recorded with a uniform and evenly spaced sampling interval. Nevertheless, it is observed that the extracted sub-dataset from the primary database may contain several missing time steps (or timestamps). In such a case, it is recommended to check for such missing timestamps by simply calculating the gradient of timestamps, and for each missing timestamp, just add an empty row consisting of only the missing timestamp value. Finally, the dataset should be sorted according to the timestamps, resulting in a uniform and evenly-spaced list of samples.

A similar procedure can be adopted for a noon report dataset. The noon reports are generally recorded every 24 h, but it may sometimes be more or less than 24 h if the vessel's local time zone is adjusted, especially on the day of arrival or departure. However, the above procedure may not be feasible in the case of AIS data, as the samples here are generally sporadically distributed. The samples in AIS data are collected at different frequencies depending on the ship's moving state, surrounding environment, traffic, and the type of AIS receiving station (land-based or satellite). It is observed here that the data is collected in short and continuous sections of the time series, leaving some significant gaps between samples, as shown in Fig. 2. Therefore, it is recommended to first resample the short and continuous sections of AIS data to a

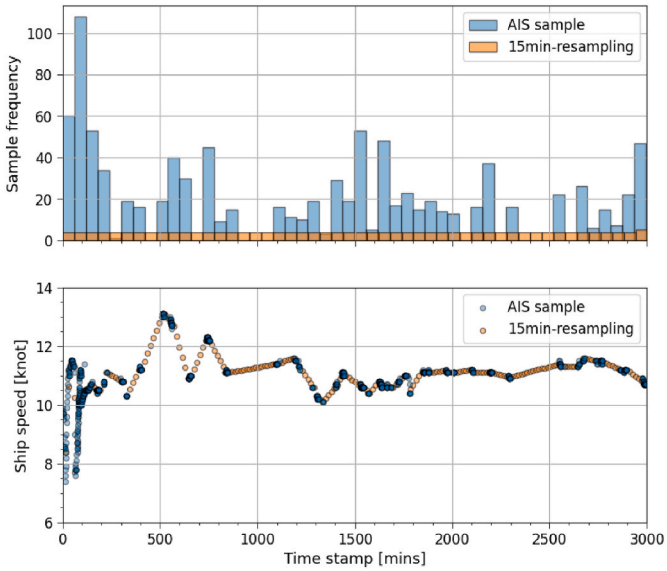
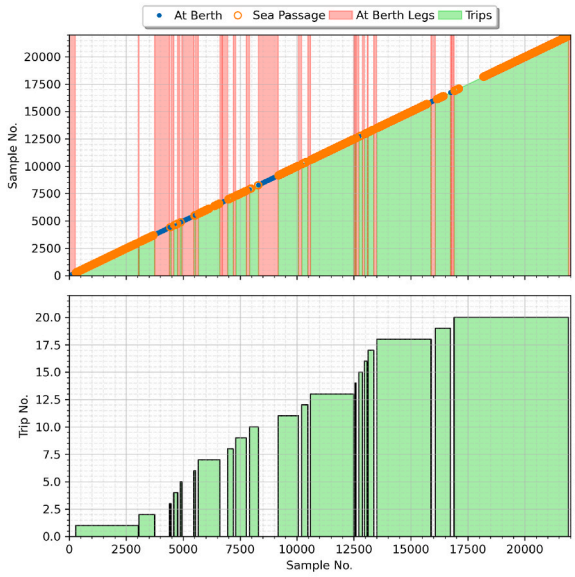
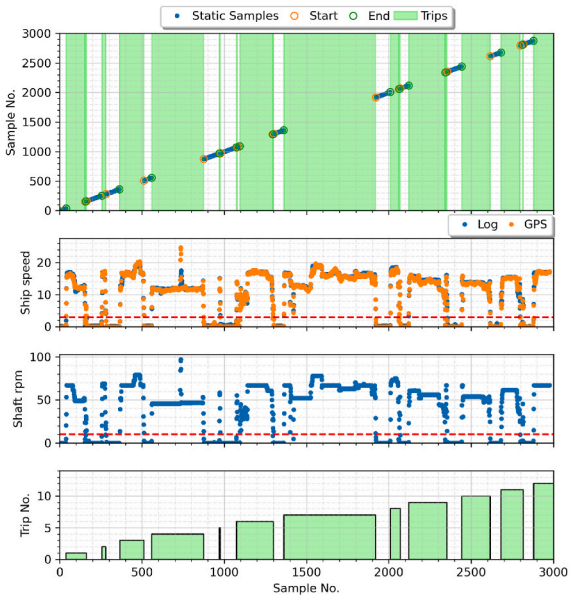


Fig. 2. Down-sampling the collected AIS data to 15-min intervals.



(a) Splitting time series into trips using the 'State' variable.



(b) Splitting time series into trips using threshold values (indicated by dashed red lines) for shaft rpm (10 rpm) and GPS speed (3 knots) variables.

Fig. 3. Splitting time series into trips.
(a) Splitting time series into trips using the 'State' variable. (b) Splitting time series into trips using threshold values (indicated by dashed red lines) for shaft rpm (10 rpm) and GPS speed (3 knots) variables.

uniform sampling interval through data resampling techniques, i.e., up-sampling or down-sampling, as demonstrated by Virtanen et al. (2020), and then, fill the remaining significant gaps with empty rows.

5.2. Divide into trips

Using conventional tools like spreadsheets, data visualization as well as handling becomes a challenge if the number of samples in the dataset is enormously large. It may simply not be practical to visualize or analyze the whole time series in a single attempt. Moreover, dividing the time series into individual trips or voyages may be considered neat and help discretize the time series into sensible sections, which may be treated individually for further data processing and analysis. Analyzing an individual trip would also give a complete overview of a port-to-port journey of the ship. Dividing the data into trips and at-berth legs would also make data processing computationally less expensive as it may be possible to ignore many samples (for further steps) where the ship is not undergoing a trip (or voyage). For such samples, it may not be necessary to interpolate the hindcast, calculate hydrostatics, calculate resistance components, etc. Lastly, identifying individual trips would also make the draft and trim correction steps easier (as discussed further).

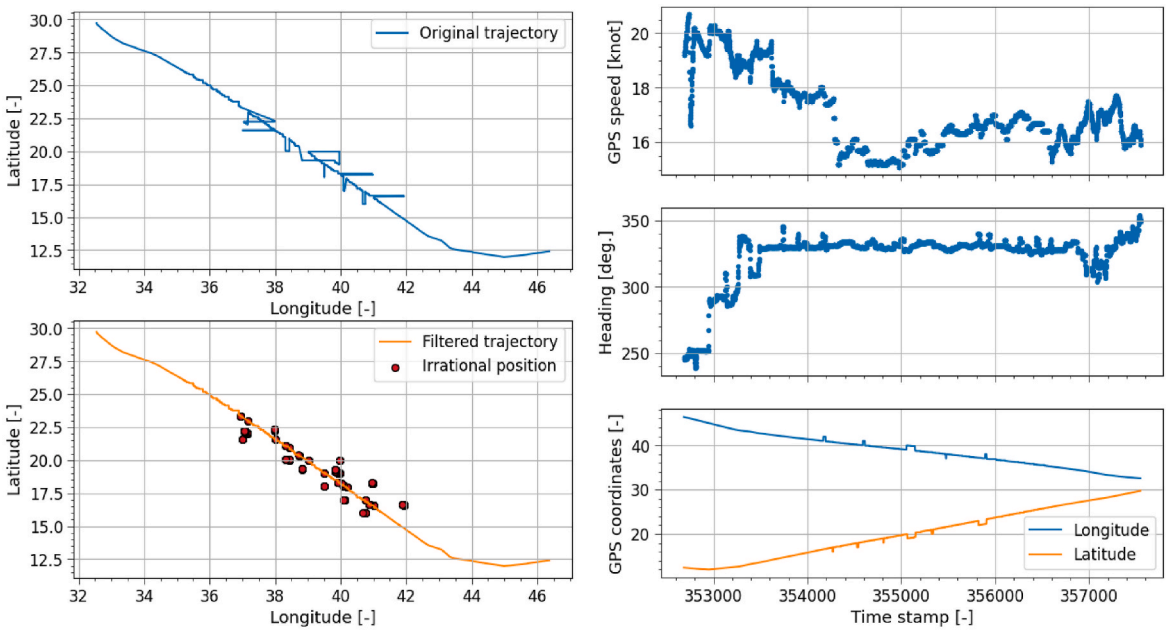
Dividing data into trips is substantially easier for noon reports and AIS data as they are generally supplied with a source and/or destination port name. In the case of in-service data, it may be possible that no such information is available. Here, if the GPS data (latitude and longitudes) is available, it may be possible to plot the samples on the world map and obtain individual trips or voyages by looking at the port calls. Alternatively, if the in-service data is supplied with a 'State' variable[4] (mentioned by Gupta et al., 2019), indicating the propulsive state of the ship, like 'Sea Passage', 'At Berth', 'Maneuvering', etc., it is recommended to find the continuous legs of 'At Berth' state and enumerate the gaps in these legs with trip numbers, containing the rest of the states, as shown in Fig. 3(a). Alternatively, it is recommended to use the shaft rpm and GPS speed (or speed-over-ground) time series to identify the starting and end of each port-to-port trip. Here, a threshold value can be adopted for the shaft rpm and GPS speed. All the samples above these threshold values (either or both) are considered in-trip samples, as shown in Fig. 3 (b). Thus, continuous legs of such in-trip samples can be enumerated with trip numbers. It may also be possible to append a few samples before and after each of these identified trips to obtain a proper trip, starting from zero and ending at zero speed and/or rpm. Such a process is designed keeping in mind the noise in the shaft rpm and GPS speed data when the ship is actually static. Finally, if the GPS data is available, further adjustments can be made by looking at the port calls on the world map plotted with the GPS data.

5.3. Interpolate hindcast & GPS position correction

Even if the raw data contains information regarding the state of the weather for each data sample, it may be an excellent idea to interpolate weather hindcast (or Metocean) data available from one of the well-established sources. The interpolated hindcast data would not only provide a quantitative measure of the weather conditions (and, consequently, the environmental loads) experienced by the ship, but it would also help carry out some necessary validation checks (discussed later in section 5.5). To interpolate hindcast data, the information regarding the location (latitude and longitude) and recording timestamp must be available in the ship's dataset. For ship performance analysis, it should be aimed that, at least, the information regarding the three main environmental load factors, i.e., wind, waves, and sea currents, is gathered from the weather hindcast sources. For a further detailed analysis, it may also be a good idea to obtain additional variables, like sea water temperature (both surface and gradient along the depth of the ship), salinity, etc.

Before interpolating the weather hindcast data to the ship's location and timestamps, it is recommended to ensure that the available GPS (or

[4] Generally available for ships equipped with Marorka systems (www.marorka.com).



(a) Original trajectory and filtered trajectory with irrational GPS position.

(b) GPS speed (or speed-over-ground), heading, and position time series for the corresponding period.

Fig. 4. GPS position correction.
(a) Original trajectory and filtered trajectory with irrational GPS position. (b) GPS speed (or speed-over-ground), heading, and position time series for the corresponding period.

navigation) data is validated and corrected (if required) for errors. If the GPS data is inaccurate, weather information at the wrong location is obtained, resulting in incorrect values for further analysis. For instance, the ship's original trajectory obtained from the GPS data, presented in Fig. 4(a), shows that the ship proceeds in a specific direction while suddenly jumping to an off-route location occasionally. The ship, of course, may have gone off-route as shown here, but referring to the GPS speed and heading of the ship at the corresponding time, shown in Fig. 4 (b), it is evident that the navigation data is incorrect. Here, such an irrational position change can be detected through the quasi-steady-state filter, explained in section 3. The 'irrational position' in Fig. 4(a) shows the coordinates identified as unsteady when the quasi-steady-state filter is applied to the longitude and latitude time series. The 'irrational position' can, then, be fixed by linearly interpolating the latitude and longitude values using the adjacent data.

The hindcast data sources generally allow downloading a subset of the variables, timestamps, and a sub-grid of latitudes and longitudes, i.e., the geographical location. Depending on the hindcast source, the datasets can be downloaded manually (by filling out a form), using an automated API or openDAP script, or even by directly accessing their FTP servers. It may also be possible to select the temporal and spatial resolution of the downloaded variables. In some cases, the hindcast web servers allow the users to send a single query, in terms of location, timestamp, and list of variables, to extract the required data for an individual data sample, generally using the openDAP interface. However, every query received by these servers is generally queued for processing, causing substantially long waiting times, as they are facing a good amount of traffic from all over the world. Thus, it is recommended to simply download the required subset of data on a local machine for faster interpolation. Once the hindcast data files are available offline, the main task is to understand the cryptic (but highly efficient) data packaging format. Nowadays, the two most popular formats for such data files are GRIdded Binary data (GRIB) and NetCDF. GRIB (available as GRIB1 or GRIB2) is the international standard accepted by World Meteorological Organization (WMO). However, due to some compatibility issues with Windows operating systems, it may be preferable to use the NetCDF format.

Finally, a step-by-step interpolation has to be carried out for each data sample from the ship's dataset. Algorithm 1 shows a simple procedure for a linear interpolation scheme. Here, the spatial and temporal interpolation is performed in steps 10 and 12, respectively. For a simple and reliable procedure, it is recommended to perform the spatial interpolation using a grid of latitudes and longitudes around the ship's location, after fitting a linear or non-linear 2D surface over the hindcast grid. It may be best to use a linear surface here as, firstly, the hindcast data may not be so accurate that performing a higher order interpolation would provide any better estimates, and secondly, in some cases, higher order interpolation may result in highly inaccurate estimates, due to the waviness of the over-fitted non-linear surface. Similar arguments can be made in the case of temporal interpolation, and therefore, linear interpolation in time can also be considered acceptable. The advantage of using the given algorithm is that the interpolation steps, here, can be easily validated by plotting contours (for spatial interpolation) and time series (for temporal interpolation). It should be noted here that this algorithm (Algorithm 1) would need some modification in case of angular measurements, as the straightforward spatial and temporal interpolation would result in an error similar to the time averaging problem discussed in section 4.5. Thus, in the case of angular measurements, the values should be transformed using the trigonometric sine and cosine transformations before interpolation, and then, transformed back into the angular measurements using *arctangent* or arctan transformation after spatial and temporal interpolation, as suggested by Grancher et al. (2012).

Algorithm 1. A simple algorithm for linear interpolation of weather hindcast data variables.

1: $wD \leftarrow$ weather hindcast data
2: $x \leftarrow$ data variables to interpolate from hindcast ▷ Also contained in wD
3: $wT \leftarrow$ all the timestamps available in wD
4: for all timestamps in ship's dataset do ▷ Iterates over i
5: $t_i \leftarrow$ current ship timestamp
6: $loc_i \leftarrow$ current ship location ▷ latitude & longitude
7: $wt \leftarrow$ 2 timestamps from wT, one just before t_i and one after ▷ t_i lies between these 2 timestamps
8: **for all** x **do** ▷ Iterates over j
9: **for all** wt **do** ▷ Iterates over k
10: $sD[x_j][t_i][wt_k] \leftarrow$ 2D (spatial) interpolation on $wD[x_j][wt_k, loc_{all}]$ at x_j and loc_i
11: **end for**
12: $tD[x_j][t_i] \leftarrow$ temporally interpolated value for x_j at t_i using $sD[x_j][t_i]$
13: **end for**
14: **end for**

An important feature of hindcast datasets is masking invalid values. For instance, the significant wave height should only be predicted by the hindcast model for the grid nodes which fall into the sea. Therefore, requesting the value of such a variable on land should result in an invalid value. Such invalid values (or nodes) are by default masked in the downloaded hindcast data files, probably for efficient data storage. These masked nodes should be filled with zeros[5] before carrying out the spatial interpolation in step 10, as one or more of these nodes may contribute to the interpolation. Alternatively, if a particular masked node is contributing to the interpolation, it can be set to the mean of other nodes surrounding the point of interpolation, as suggested by Ejdfors (2019). It is argued by Ejdfors (2019) that this would help avoid artificially low (zero) values during the interpolation. However, calculating the mean for all the required nodes may be computationally expensive. Therefore, substituting the masked node with the nearest node value before carrying out the interpolation or just using the nearest node value for the point of interpolation, which is surrounded by at least 1 masked node, may be much more efficient.

5.4. Derive new features

Interpolating the weather hindcast variables to the ship's location at a given time would provide the hindcast variables in the global (or the hindcast model's) reference frame. For further analysis, it may be appropriate to translate these variables to the ship's frame of reference, and furthermore, it may be desired to calculate some new variables which could be more relevant for the analysis or could help validate the assimilated (ship and hindcast) dataset. The wind and sea current variables can be resolved into the longitudinal and transverse speed components for validation and further analysis. Unfortunately, the wave load variables cannot be resolved in a similar manner, but the mean wave direction should be translated into the relative mean wave direction (relative to the ship's heading or course).

5.5. Validation checks

Although it is recommended to validate each processing step through visualizations (or plots), it may be a good idea to take an intermediate pause and perform all types of possible validation checks. These validation checks would not only help assess the dataset from the reliability point of view but can also be used to understand the correlation between various features. The validation checks can be done top-down, starting from the most critical feature to the least one. As explained in section 2.4, the shaft power measurements can be validated against the shaft rpm and shaft torque measurements, if these are available, else just plotting the shaft rpm against the shaft power can also provide a good insight into the quality of data. For a better assessment, it is suggested to visualize the shaft rpm vs shaft power overlaid with the engine operational envelope and propeller curves, as presented by Liu et al. (2020a) (in figure 11). Any sample falling outside the shaft power overload line (especially at high shaft rpm) should be removed from the analysis, as they may be having measurement errors. It may also be possible to make corrections (or remove affected samples) if the shaft power data seems to be shifted (up or down) with respect to the propeller curves due to sensor bias or mechanical failure related to the propeller and/or shaft bearings.

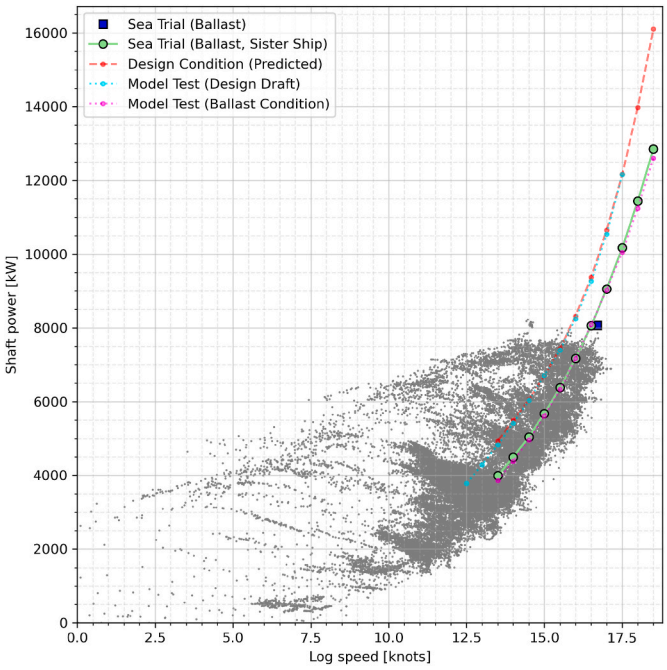


Fig. 5. Speed-through-water (log speed) vs shaft power with various estimates of speed-power calm-water curves.

The quality of speed-through-water measurements can be assessed by validating it against its estimate, obtained as a difference between the speed-over-ground and longitudinal current speed. Here, it should be kept in mind that the two sets of values may not be a very good match due to several problems, discussed in section 4.3. Although the speed-over-ground, measured using the onboard GPS sensor, can be pretty accurate, the current or sea water speed, which is seldom recorded onboard, tells an entirely different story. The sea water speed, generally obtained from hindcast sources, is not accurate enough to obtain a good estimate for speed-through-water, as indicated by Antola et al. (2017). It should also be noted that the temporal and spatial resolution of weather hindcast data is relatively larger than the sampling interval of the in-service data recorded onboard the ship. Moreover, the sea water speed or sea currents vary along the depth of the sea. Therefore, the incident longitudinal sea water speed must be calculated as an integral of the sea water speed profile over the ship's depth. Thus, to obtain accurate estimates for the speed-through-water, the sea water speed has to be measured or estimated up to a certain depth of the sea with good enough accuracy, which is not possible with the current state-of-the-art methods. Nevertheless, visualizing the speed-through-water vs shaft power along with all the available estimates of the speed-power calm-water curve is an important validation step (shown in Fig. 5). Here, the majority of measurement data should accumulate around these curves. In case of disparity between the curves, the curve obtained through the sea trial of the actual ship may take precedence.

The interpolated weather hindcast data variables must also be validated against the measurements taken onboard the ship. This is quite critical as the sign and direction notations assumed by the hindcast

[5] This can be done easily in python using numpy.ma.filled.

models and the ship's sensors (or data acquisition system) are probably not the same, which may cause mistakes during the interpolation step. Moreover, most ships are generally equipped with anemometers that can measure the actual and relative wind speed and directions. These two modes (actual or relative) can be switched through a simple manipulation by the crew onboard. It is possible that this mode change may have occurred during the data recording duration, resulting in errors in the recorded data. In addition, there may be a difference between the reference height of the wind hindcast data and the vertical position of the installed anemometer, which may lead to somewhat different results even at the same location at sea. The wind speed at the reference height (V_{WTref}) can be corrected using the anemometer recorded wind speed (V_{WT}), assuming a wind speed profile, as follows (recommended by ITTC, 2021):

$$V_{WT\,ref} = V_{WT}\left(\frac{Z_{ref}}{Z_a}\right)^{\frac{1}{9}} \tag{3}$$

where Z_{ref} is the reference height above the sea level (generally assumed 10 m, which is also adopted for most of the hindcast models) and Z_a is the height of the anemometer.

Finally, these wind measurements can be translated into the longitudinal and transverse relative components. The obtained transverse relative wind speed can be validated against the transverse wind speed, obtained from the hindcast source, as they are more or less the same. Similarly, the difference between the longitudinal relative wind speed and the speed-over-ground of the ship can be validated against the longitudinal wind speed obtained from hindcast, as shown in Fig. 6. In the case of time-averaged in-service data, the problem of faulty averaging of angular measurements when the measurement values are near 0 or 360° (i.e., the angular limits), explained in section 4.5, must also be verified and appropriate corrective measures should be taken. From Fig. 6, it can be clearly seen that the time-averaging problem (in relative wind direction) causes the longitudinal wind speed (estimated using the ship data) to jump from positive to negative, resulting in a mismatch

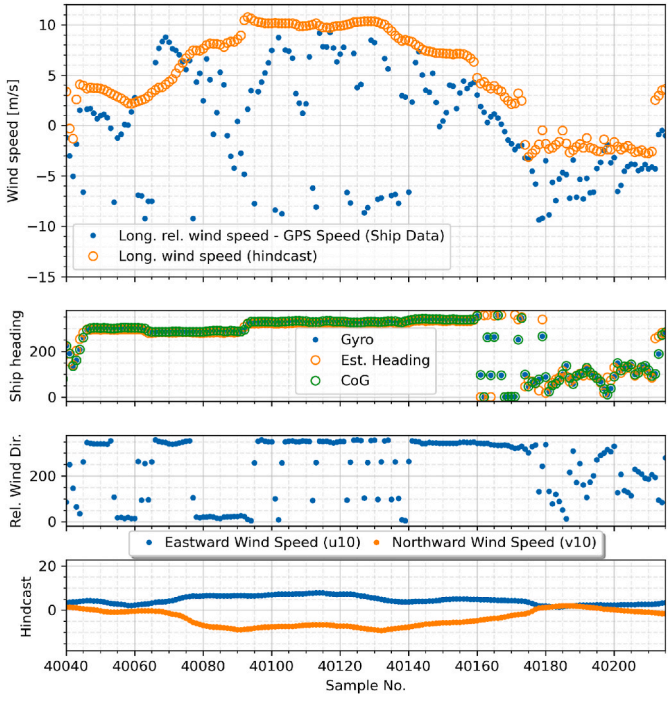


Fig. 6. Validating longitudinal wind speed obtained using the ship data against the values obtained from the hindcast. The time-averaging problem with angular measurements around 0 or 360° (explained in section 4.5) is clearly visible here.

with the corresponding hindcast values. In such a case, it is recommended to either fix these faulty measurements, which may be difficult as there is no proven way to do it or just use the hindcast measurements for further analysis.

As discussed in the case of noon reports in section 4.6, weather information generally refers to the state of the weather at the time when the report is logged, which is probably not the average state from noon to noon. Furthermore, the wind loads here are observed based on the Beaufort scale. Therefore, the deviation may be somewhat large when converted to the velocity scale. In this case, it is recommended to consider the daily average values obtained from the weather hindcast data, over the travel region, rather than the noon report values.

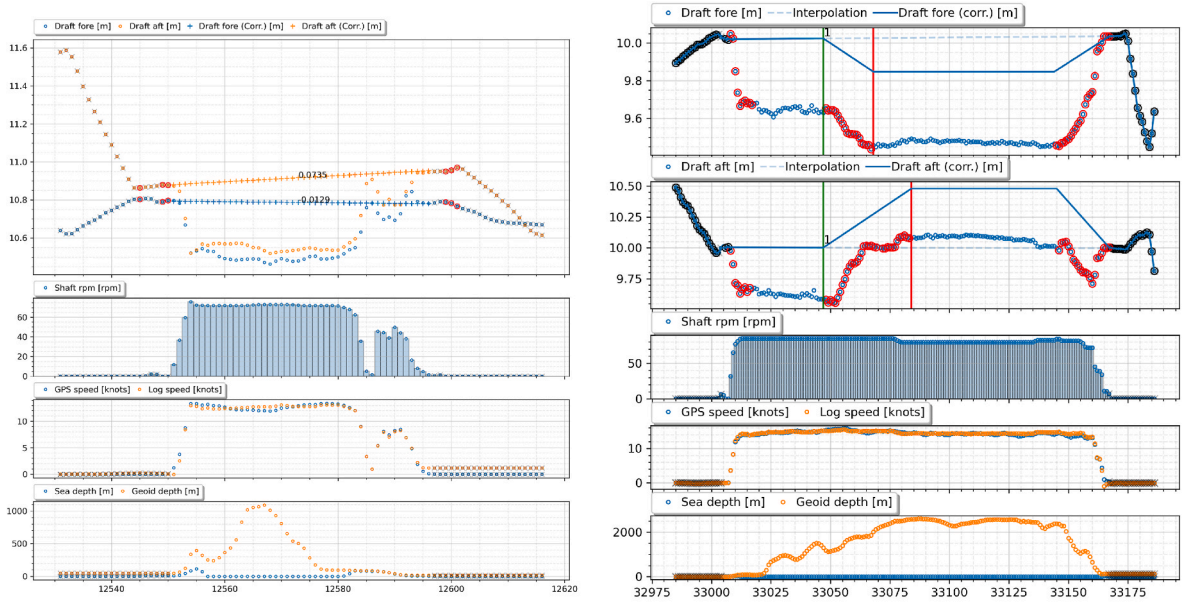
5.6. Data processing errors

The validation step is very critical in finding out any processing mistakes or inherent problems with the dataset, as demonstrated in the previous section. Such problems or mistakes, if detected, must be corrected or amended before moving forward with the processing and analysis. The main mistakes found at this step are generally either interpolation mistakes or incorrect formulation of the newly derived feature. These mistakes should be rectified accordingly, and the data processing should continue further, as shown in the flow diagram (Fig. 1).

5.7. Fix draft & trim

The draft measurements recorded onboard the ship are often found to be incorrect due to the Venturi effect, explained briefly in section 4.3. The Venturi effect causes the draft measurements to drop to a lower value due to a non-zero negative dynamic pressure as soon as the ship develops a relative velocity with respect to the water around the hull. It may seem like a simple case, and one may argue that the measurements can be fixed by just adding the water level height equivalent to the hydrodynamic pressure, which may be calculated using the ship's speed-through-water. Here, it should be noted that, firstly, to accurately calculate the hydrodynamic pressure, one would need the localized relative velocity of the flow (and not the ship's speed-through-water), which is impractical to measure. Secondly, the speed-through-water measurements are also known to have several sources of inaccuracy, as discussed previously in sections 4.3 and 5.5. Alternatively, it may be possible to obtain the correct draft measurements from the ship's loading computer. The loading computer can calculate the draft and trim in real-time based on information such as the ship's lightweight, cargo weight and distribution, and ballast water loading configuration. However, as per the usual practice, the way to fix these incorrect measurements is by interpolating the draft during a voyage using the draft measured just before and after the voyage. Such a simple solution provides good results for a simple case where the draft of the ship basically remains unchanged during the voyage, except for the reduction of the draft due to consumed fuel, as shown in Fig. 7(a).

In a more complex case where the draft of the ship is changed in the middle of the voyage and the ship is still moving, i.e., conducting ballasting operations or trim adjustments during transit, the simple draft interpolation would result in corrections which can be way off the actual draft of the vessel. As shown in Fig. 7(b), the fore draft is seen to be dropping and the aft draft increasing in the middle of the voyage without much change in the vessel speed, indicating trim adjustments during transit. In this case, a more complex correction can be applied after taking into account the change in the draft during the transit. Here, first of all, a draft change operation is identified (marked by green and red vertical lines in Fig. 7(b)), then the difference between the measurements before and after the operation is calculated by taking an average over a number of samples. Finally, a ramp is created between the start (green line) and end (red line) of the draft change operation. The slope of the ramp is calculated using the difference between the draft



(a) Simple draft correction.

(b) Complex draft correction.

Fig. 7. Correcting in-service measured draft.
(a) Simple draft correction. (b) Complex draft correction.

measurements before and after the draft change operation. The draft change operation can either be identified manually, by looking at the time series plots, or by using the first stage of the quasi-steady-state filter (presented in section 3) with not-so-strict settings. The latter is employed in the case presented in Fig. 7(b).

In the case of AIS data, Bailey et al. (2008) reported that 31% of the draft information out of the investigated AIS messages had obvious errors. The draft information from AIS data generally corresponds to the condition of ships while arriving at or departing from the port, and changes due to fuel consumption and ballast adjustment onboard are rarely updated. Since the draft obtained from the AIS as well as noon reports has a long update cycle and is acquired manually, it is practically difficult to precisely fix the draft values as in the case of in-service data. However, by comparing the obtained draft with a reference value, it may be possible to gauge whether the obtained draft is, in fact, correct. If the obtained draft excessively deviates from the reference, it may be possible to remove the corresponding data samples from further analysis or replace the obtained draft value with a more appropriate value. Table 2 shows the results of investigating the average draft ratio, which is the ratio of the actual draft (T_c) and design draft (T_d), for various ship types from 2013 to 2015 by Olmer et al. (2017). As summarized in the table, the draft ratio varies depending on the ship type and the voyage type. Using these values as the above-mentioned reference, the draft obtained from the AIS data and noon reports can be roughly checked and corrected.

Table 2
Average draft ratio (T_c/T_d) for different ship types. T_c = actual draft during a voyage; T_d = design draft of the ship.

Ship types	Ballast Voyage	Laden Voyage
Liquefied gas tanker	0.67	0.89
Chemical tanker	0.66	0.88
Oil tanker	0.60	0.89
Bulk carrier	0.58	0.91
General cargo	0.65	0.89
The following ship types do not generally have ballast-only voyages.		
Container	0.82	
Ro-Ro	0.87	
Cruise	0.98	
Ferry pax	0.90	
Ferry ro-pax	0.93	

Source: Olmer et al. (2017).

5.8. Calculate hydrostatics

Depending on the type of performance analysis, it may be necessary to have features like displacement, Wetted Surface Area (WSA), etc., in

Table 3
Estimation formulas for wetted surface area (WSA) of different ship types.

Category	Formula	Reference
Tanker/Bulk carrier	$WSA = 0.99 \cdot (\frac{\nabla}{T} + 1.9 \cdot L_{WL} \cdot T)$	Kristensen and Bingham (2017)
Container	$WSA = 0.995 \cdot (\frac{\nabla}{T} + 1.9 \cdot L_{WL} \cdot T)$	Kristensen and Bingham (2017)
Other (General)	$WSA = 1.025 \cdot (\frac{\nabla}{T} + 1.7 \cdot L_{PP} \cdot T)$	Molland (2011)

Table 4
Typical block coefficient (C_B) range at design draft for different ship types, given by MAN Energy Solutions (2018).

Category	Type	Block coefficient (C_B)
Tanker	Crude oil carrier	0.78–0.83
	Gas tanker/LNG carrier	0.65–0.75
	Product	0.75–0.80
	Chemical	0.70–0.78
Bulk carrier	Ore carrier	0.80–0.85
	Regular	0.75–0.85
Container	Line carrier	0.62–0.72
	Feeder	0.60–0.70
General cargo	General cargo/Coaster	0.70–0.85
Roll-on/roll-off cargo	Ro-Ro cargo	0.55–0.70
	Ro-pax	0.50–0.70
Passenger ship	Cruise ship	0.60–0.70
	Ferry	0.50–0.70

the dataset, as they are more relevant from a hydrodynamic point of view. Moreover, most of the empirical or physics-based methods for resistance calculations (to be done in the next step) requires these features. Unfortunately, these features cannot be directly recorded onboard the ship. However, it is reasonably convenient to estimate them using the ship's hydrostatic table or hull form (or offset table) for the corresponding mean draft and trim for each data sample. Here, it is recommended to use the corrected draft and trim values, obtained in the previous step. If the detailed hull form is not available, the Wetted Surface Area (WSA) can also be estimated using the empirical formulas shown in Table 3. The displacement at the design draft, on the other hand, can be estimated using the ship's principle particulars and typical range of block coefficient (C_B), presented in Table 4.

5.9. Calculate resistance components

There are several components of the ship's total resistance, and there are several methods to estimate each of these components. The majority of a ship's total resistance comprises three main components: calm-water, added wind, and added wave resistance. It is possible to further divide the calm-water resistance into sub-components, namely, skin friction and residual resistance.

Calm-water resistance. The total calm-water resistance can be calculated using one of the many well-known empirical methods, like Guldhammer and Harvald (1970), updated Guldhammer and Harvald (Kristensen and Bingham, 2017), Hollenbach (1998), Holtrop and Mennen (1982), etc. These empirical methods are developed using the data from numerous model test results of different types of ships, and each one is proven to be fitting well on several different ship types. The latter makes choosing the right method for a ship quite complicated. The easiest way to select the proper calm-water resistance estimation method is to calculate the calm-water resistance from each method and compare it with the corresponding data obtained for the given ship. The calm-water data for a given ship can be obtained from the model tests, sea trials, CFD analyses, or even after filtering the operational data for near-calm-water conditions from the new-built ship. The usual practice here is to use the sea trial data as it is obtained and corrected for near-calm-water conditions and does not suffer from scale effects, as seen in model test results. However, the sea trials are sometimes conducted at only the high-speed range and ballast displacement (as shown in Fig. 5). Thus, it is recommended to use the near-calm-water filtered (and corrected) operational data (in new-built or negligible fouling condition) or thoroughly validated full-scale CFD results for selecting the suitable method, so that a good fit can be ensured for a complete range of speed and displacement.

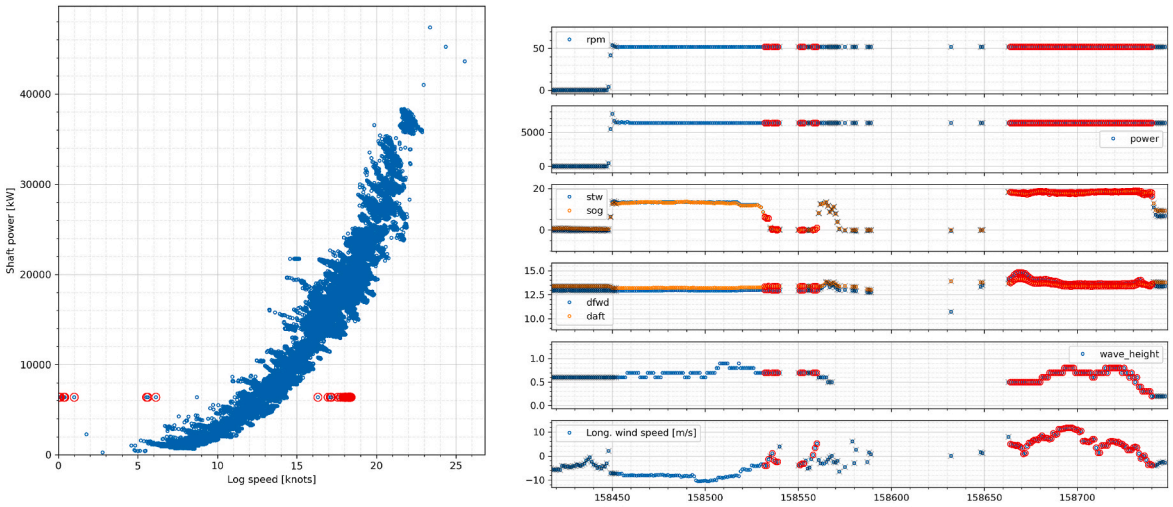
Added wind resistance. According to ITTC (2021), the increase in resistance due to wind loads can be obtained by applying one of the three suggested methods, namely, wind tunnel model tests, STA-JIP, and Fujiwara's method. If the wind tunnel model test results for the vessel are available, it may be considered the most accurate method for estimating added wind resistance. Otherwise, the database of wind resistance coefficients established by STA-JIP (van den Boom et al., 2013) or the regression formula presented by Fujiwara et al. (2005) is recommended. From the STA-JIP database, experimental values according to the specific ship type can be obtained. In contrast, Fujiwara's method is based on the regression analysis of data obtained from several wind tunnel model tests for different ship types. The two main sets of parameters required to estimate the added wind resistance using the above three methods are incident wind parameters and information regarding the exposed area to the wind. The incident wind parameters, i.e., relative wind speed and direction, can be obtained from onboard measurements or weather hindcast data. In the case of weather hindcast data, the relative wind measurements can be calculated from the interpolated hindcast values according to the formulation outlined by ITTC (2021) in section E.1. In the case of onboard measurements, the relative wind measurements should be corrected for the vertical position of the anemometer according to the instructions given by ITTC (2021) in section E.2, also explained here in section 5.5. The information regarding the exposed area to the wind (with the varying draft or loading condition) can be either estimated using the general arrangement drawing of the ship or approximately obtained using a regression formula based on the data from several ships, presented by Kitamura et al. (2017).

Added wave resistance. The added wave resistance (R_{AW}) can also be obtained similarly using one of the several well-established methods. ITTC (2021) recommends conducting seakeeping model tests in regular waves to get R_{AW} transfer functions, which can further be used to estimate R_{AW} for the ship in irregular seas. Alternatively, it is recommended to obtain R_{AW} using a physics-based empirical method like STAWAVE1 and STAWAVE2. STAWAVE1 is a simplified method for directly estimating R_{AW} in head wave conditions only, and it requires limited input, including the ship's waterline length, breadth, and significant wave height. STAWAVE2 is an advanced method to empirically estimate parametric R_{AW} transfer functions for a ship. As presented by van den Boom et al. (2013), STAWAVE1 is applicable to only short waves (compared to the ship length and speed), whereas STAWAVE2 is applicable to long swells, when the resistance due to ship motions also becomes important. STAWAVE2 is developed using an extensive database of seakeeping model test results from numerous ships. Still, unfortunately, it only provides transfer functions for approximate head wave conditions (0 to $\pm 45^\circ$ from the bow). A method proposed by DTU (Martinsen, 2016; Taskar and Andersen, 2019, 2021) provides transfer functions for head-to-beam seas, i.e., 0 to $\pm 90^\circ$ from the bow. Finally, for all wave headings, it is possible to use one of these recently developed methods: (a) SNNM (SHOPERA-NTUA-NTU-MARIC) method, proposed and validated by Liu et al. (2020b) and Wang et al. (2021), respectively; (b) CTH (Chalmers Tekniska Högskola) method, proposed by Lang and Mao (2021); and (c) A method combining (a) and (b), proposed by Kim et al. (2022). Based on the validation results from Wang et al. (2021), ITTC (2021) has recommended the SNNM method (Liu et al., 2020b) for correcting speed/power trial data. However, the guidelines from ITTC (2021) are only applicable for mild weather conditions, necessary during the speed/power trials. Moreover, the findings presented by Kim et al. (2022) suggest that the method combining the SNNM and CTH methods may result in improved accuracy, particularly for fleet-level analyses encompassing diverse ship characteristics and operating conditions. Thus, there may still be some scope for improvement.

5.10. Data cleaning & outlier detection

It may be argued by some that the process of data cleaning and outlier detection should be carried-out way earlier in the data processing framework, as proposed by Dalheim and Steen (2020a), but it should be noted that all the above steps presented here have to be performed only once for a given dataset, whereas data cleaning is done based on the features (or variables) selected for further analysis. Since the same dataset can be used for several different analyses, each of which may be using different sets of features, some part of data cleaning has to be repeated before each analysis to obtain a clean dataset with as many data samples as possible. Moreover, the additional features acquired during the above-listed processing steps may be helpful in determining to a better extent if a suspected sample is actually an outlier or not.

It may be possible to reduce the workload for the above processing steps by performing some basic data cleaning before some of these steps. For instance, while calculating the resistance components for in-trip data samples, it is possible to filter out samples with invalid values for one or more of the data variables used to calculate these components, like speed-through-water, mean draft (or displacement), etc. This would reduce the number of samples for which the new feature has to be calculated. It should also be noted that even if such simple data cleaning (before each step) is not performed, these invalid samples would be



(a) Log speed or speed-through-water (stw) vs shaft power.

(b) Time series showing the highlighted (in red) samples.

Fig. 8. Correlation-defying outliers marked with red circles.
(a) Log speed or speed-through-water (stw) vs shaft power. (b) Time series showing the highlighted (in red) samples.

easily filtered-out in the present step. Thus, the reliability and efficacy of the data processing framework are not affected by performing the data cleaning and outlier detection step at the end.

Quasi-steady assumption. Most of the methods developed for ship performance monitoring assume that the ship is in a quasi-steady-state for each data sample. The quasi-steady-state assumption, also explained in section 3, indicates that the propulsive state of the ship remains more or less constant during the sample recording duration, i.e., the ship is neither accelerating nor decelerating, and it is not changing its direction or heading. This is especially critical for the onboard recorded time-averaged values, as the averaging duration can be substantially longer, generally, up to 15 min, hiding the effects of acceleration, deceleration, and change in vessel heading. Here, the two-stage quasi-steady-state filter, explained in section 3, can be applied to the shaft rpm and vessel heading time series to remove the samples associated with these unsteady transitions, leaving out only quasi-steady samples for further analysis. In tandem with the quasi-steady-state filter on the shaft rpm time series, it may also be possible to use the quasi-steady-state filter, with a relaxed setting, on the speed-over-ground (or any other variable) time series to filter out the samples where the signal from the sensor suddenly drops or recovers from a dead state, resulting in measurement errors. However, a strict quasi-steady-state filter applied to the ship's speed (speed-through-water or speed-over-ground), torque, or power would result in removing most of the samples that are influenced by the environmental loads, which is certainly not desired in most cases. Thus, caution must be used while using such tools in practice.

Outliers. As discussed in section 4.4, the outliers can be divided into two broad categories: (a) Contextual outliers, and (b) Correlation-defying outliers. The contextual outliers can be identified and resolved by the methods presented as well as demonstrated by Dalheim and Steen (2020a), and for correlation-defying outliers, methods like Principal Component Analysis (PCA) and autoencoders can be used. However, no good reference could be obtained for the latter from the reviewed literature, presenting a good scope for some future work. Fig. 8 shows the in-service data samples recorded onboard a ship. The data here is already filtered-out for quasi-steady assumption (explained above) and contextual outliers, according to the methods suggested by Dalheim and Steen (2020a). Thus, the samples highlighted by red circles (around 6.4 MW shaft power in Fig. 8(a)) can be classified as correlation-defying outliers. The time series plot (shown in Fig. 8(b)) indicates that the

Table 5
Categorized list of variables recorded onboard the ship, and newly added variables (bottom part) created while processing the data using the presented framework.

Navigation	Auxiliary Power System	Propulsion System	Environment
Recorded Onboard			
Latitude	Aux. Consumed	State	Relative Wind Speed
Longitude	Aux. Electrical Power Output	ME Load Measured	Relative Wind Direction
Gyro Heading	DG1 Power	Shaft Power	Sea Depth
COG Heading	DG2 Power	Shaft rpm	
	DG3 Power	Shaft Torque	
		ME Consumed	
		Draft Fore	
		Draft Aft	
		GPS Speed	
		Log Speed	
		Cargo Weight	
Newly Added			
Trip No.		Mean Draft (corrected)	Wind Speed (True)
		Trim-by-aft (corrected)	Wind Direction (True)
		Displacement	Long. & Trans. Wind Speed
		Wetted Surface Area	Significant Wave Height
		Calm-water Resistance	Mean Wave Period
		Wind Resistance	Mean Wave Direction
		Added Wave Resistance	Current Speed
			Current Direction
			Long. Current Speed

Abbreviations: IMO = International Maritime Organization; COG = Center of Gravity; Aux. = Auxiliary; DG = Diesel Generator (for auxiliary power systems); ME = Main Engine (for propulsion system); GPS = Global Positioning System; Long. = Longitudinal; Trans. = Transverse.

detected outliers have inaccurate measurements for the speed-through-water (stw) and speed-over-ground (sog), defying the correlation between these variables and the rest. It is also quite surprising to notice that the same fault occurs in both speed measurements simultaneously, considering that they are probably obtained from different sensors. Nevertheless, the time series presents a simple case of the signal dropping out and then recovering from dead, and as discussed in the previous paragraph, it may be detected using the quasi-steady-state filter with relaxed settings. However, it may be appropriate to use a more capable tool like PCA or autoencoders to carry out an in-depth correlation-defying outlier detection.

5.11. Processed data

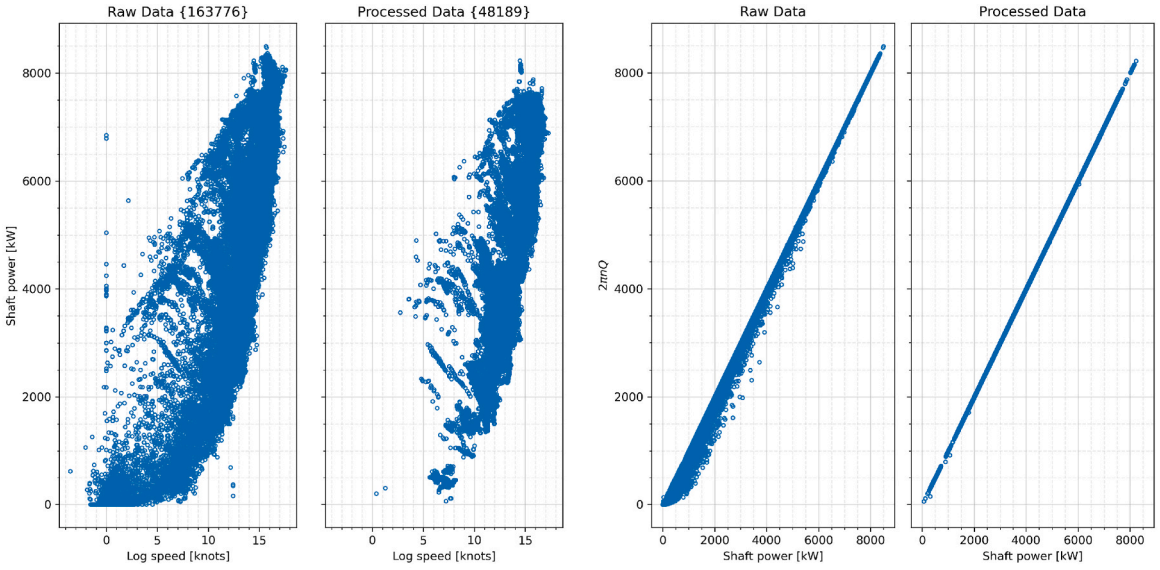
As mentioned earlier, validating the data processing framework is crucial so that reliable results can be obtained using the data processed by the framework. A thorough validation requires comparison with a benchmark or an already established standard. However, no such standard is available to validate the data processing framework. Nevertheless, it is possible to get an overview regarding the quantitative and qualitative improvements in the processed data. For instance, the most significant quantitative improvement is achieved by adding several new variables through this framework, shown in Table 5, which makes the data more readily available for the ship's hydrodynamic performance analysis. For a different type of analysis, another relevant set of variables can be added. Moreover, dividing the data into individual trips (explained in section 5.2) makes it easier to handle and visualize long time series as well as help remove (if required) the samples when the ship is static, say, at berth or anchorage. Regarding qualitative improvements, it is clearly visible from the comparison of raw and processed data, presented in Fig. 9, that the noise in the processed data is significantly reduced.

6. Conclusion

Data quality is paramount in estimating the performance of a ship. In this study, a streamlined semi-automatic data processing framework is developed to process data from multiple sources, like onboard recorded in-service data, AIS data, and noon reports in the context of ship performance analysis. The main conclusions of the current work can be listed as follows:

- The paper presented a brief overview of the generic and specific challenges associated with the aforementioned data sources and state-of-the-art in dealing with them.
- It is recommended to use onboard recorded in-service data over the other data sources for ship performance monitoring. It is considered more reliable due to its consistent and high sampling rate.
- It is observed that the AIS data and noon reports lack some critical variables required for ship performance analysis. They are also susceptible to human error, as the ship's crew manually logs some data variables recorded here.
- The proposed data processing framework demonstrated its capability to address most challenges associated with overwhelmingly large time series data obtained from the ship-in-service.

In addition to the above contributions, several extensions and improvements resulting from the current work are also identified, e.g., the incorporation of supplementary weather information enabled better estimation of environmental loads experienced by the ship. A simple algorithm was presented to effectively interpolate the hindcast data to the ship's location at a given time. The draft measurements recorded onboard the ships are known to have errors due to the so-called Venturi effect, generally fixed using simple linear interpolation over a voyage. If the draft or trim is voluntarily adjusted during the voyage, the simple interpolation technique fails, therefore, an ad hoc method is suggested. The inaccuracies observed in the speed-through-water measurements still stand unaddressed.



(a) Raw (left) and processed (right) in-service data from a ship. The numbers in curly brackets ({}) in the title of each subplot presents the number of samples in the corresponding subplot.

(b) Shaft power validation (comparing measured shaft power with its estimate, obtained using the measured rpm and torque) before (left) and after (right) processing the data.

Fig. 9. Comparing raw and processed (using the given framework) in-service data for a ship.
(a) Raw (left) and processed (right) in-service data from a ship. The numbers in curly brackets ({}) in the title of each subplot presents the number of samples in the corresponding subplot. (b) Shaft power validation (comparing measured shaft power with its estimate, obtained using the measured rpm and torque) before (left) and after (right) processing the data.

Estimating the resistance components can also be necessary for ship performance analysis, but choosing an appropriate method to calculate each component is critical. Therefore, it is strongly suggested to conduct validation checks to find the most suitable ways before adopting them. Such validation checks should be done, wherever possible, using the data obtained from the ship while in-service rather than just using the sea trial or model test results. Data cleaning and outlier detection are also necessary steps for processing the data. Based on some previously published literature, an improved quasi-steady-state filter is presented and found immensely useful while cleaning the data. Since cleaning the data requires selecting a subset of features (or variables) relevant to the analysis, it is recommended to perform data cleaning as the last step of the data processing framework. Some parts of it should be repeated every time before carrying out a new type of analysis. Moreover, a correlation-based outlier detection tool, like PCA or autoencoders, can be used to perform in-depth outlier detection.

The presented data processing framework processes datasets from ships-in-service systematically and efficiently, making them ready for ship performance analysis. Various data processing methods or steps mentioned here can also be used elsewhere to process the time series data from ships or similar sources, which can be used further for various tasks. Moreover, the data processing framework can be a building block for future technologies, like digital twins.

Acknowledgements

This study is part of the research projects SFI Smart Maritime - Norwegian Centre for Improved Energy-Efficiency and Reduced Emissions from the Maritime Sector (Research Council of Norway or RCN project number 237917) and CLIMMS - Climate Change Mitigation in the Maritime Sector (RCN project number 294771).

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.

References

Aldous, L., Smith, T., Bucknall, R., Thompson, P., 2015. Uncertainty analysis in ship performance monitoring. Ocean Eng. 110, 29–38.

American Bureau of Shipping, 2020. Guide for Smart Functions for Marine Vessels and Offshore Units. ABS Guides.

Antola, M., Solonen, A., Pyorre, J., 2017. Notorious speed through water. In: 2nd Hull Performance & Insight Conference (HullPIC'17), pp. 156–165.

Bailey, N.J., Ellis, N., Sampson, H., 2008. Training and Technology Onboard Ship: How Seafarers Learned to Use the Shipboard Automatic Identification System (AIS), Seafarers International Research Centre (SIRC). Cardiff University.

Carchen, A., Atlar, M., 2020. Four KPIs for the assessment of biofouling effect on ship performance. Ocean Eng. 217, 107971 https://doi.org/10.1016/j.oceaneng.2020.107971. URL: https://www.sciencedirect.com/science/article/pii/S0029801820309239.

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 Eng. 186 https://doi.org/10.1016/j.oceaneng.2019.05.045 cited By 1.

Dalheim, Ø.Ø., Steen, S., 2020a. Preparation of in-service measurement data for ship operation and performance analysis. Ocean Eng. 212, 107730 https://doi.org/10.1016/j.oceaneng.2020.107730. URL: https://www.sciencedirect.com/science/article/pii/S0029801820307125.

Dalheim, Ø., Steen, S., 2020b. A computationally efficient method for identification of steady state in time series data from ship monitoring. J. Ocean Eng. Sci. https://doi.org/10.1016/j.joes.2020.01.003 cited By 1.

Dalheim, Ø.Ø., Steen, S., 2021. Uncertainty in the real-time estimation of ship speed through water. Ocean Eng. 235, 109423 https://doi.org/10.1016/j.oceaneng.2021.109423. URL: https://www.sciencedirect.com/science/article/pii/S0029801821008313.

Ejdfors, K.O., 2019. Use of In-Service Data to Determine the Added Power of a Ship Due to Fouling. Master's thesis, NTNU. URL: http://hdl.handle.net/11250/2622960.

Farag, Y.B., Ölçer, A.I., 2020. The development of a ship performance model in varying operating conditions based on ann and regression techniques. Ocean Eng. 198, 106972 https://doi.org/10.1016/j.oceaneng.2020.106972. URL: https://www.sciencedirect.com/science/article/pii/S0029801820300536.

Foteinos, M., Tzanos, E., Kyrtatos, N., 2017. Ship hull fouling estimation using shipboard measurements, models for resistance components, and shaft torque calculation using engine model. J. Ship Res. 61, 64–74. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85024491367&doi=10.5957%2fJOSR.61.2.160053&partnerID=40&md5=4d8fedda376f0ed57cf089dbbfe28012. doi:10.5957/JOSR.61.2.160053, cited By 3.

Fujiwara, T., Ueno, M., Ikeda, Y., 2005. A new estimation method of wind forces and moments acting on ships on the basis of physical component models. J. Jpn. Soc. Nav. Archit. Ocean Eng. 2, 243–255. https://doi.org/10.2534/jjasnaoe.2.243.

Grancher, D., Bar-Hen, A., Paris, R., Lavigne, F., Brunstein, D., 2012. Spatial interpolation of circular data: Application to tsunami of December 2004. Adv. Appl. Stat. 30.

Guldhammer, H., Harvald, S., 1970. Ship Resistance: Effect of Form and Principal Dimensions. Akademisk Forlag.

Guo, B., Gupta, P., Steen, S., Tvete, H.A., 2023. Evaluating vessel technical performance index using physics-based and data-driven approach. Ocean Eng. 286, 115402 https://doi.org/10.1016/j.oceaneng.2023.115402. URL: https://www.sciencedirect.com/science/article/pii/S0029801823017869.

Gupta, P., Rasheed, A., Steen, S., 2022. Ship performance monitoring using machine-learning. Ocean Engineering 254, 111094.

Gupta, P., Steen, S., Rasheed, A., 2019. Big data analytics as a tool to monitor hydrodynamic performance of a ship. In: Ocean Engineering of International Conference on Offshore Mechanics and Arctic Engineering, ume vol. 7A. https://doi.org/10.1115/OMAE2019-95815,v07AT06A059.

Gupta, P., Taskar, B., Steen, S., Rasheed, A., 2021. Statistical modeling of ship's hydrodynamic performance indicator. Appl. Ocean Res. 111, 102623 https://doi.org/10.1016/j.apor.2021.102623. URL: https://www.sciencedirect.com/science/article/pii/S0141118721001000.

Hansen, S.V., Petersen, J., Jensen, J., 2011. Performance Monitoring of Ships. DTU Mechanical Engineering.

Hollenbach, K.U., 1998. Estimating resistance and propulsion for single-screw and twin-screw ships. Ship Technol. Res. 45, 72–76. Cited By 21.

Holtrop, J., Mennen, G., 1982. Approximate Power Prediction Method, vol. 29, pp. 166–170. https://doi.org/10.3233/isp-1982-2933501 cited By 347.

ISO, ISO 19030-2, 2016. Ships and Marine Technology — Measurement of Changes in Hull and Propeller Performance — Part 2: Default Method, 2016. URL: https://www.iso.org/standard/63775.html.

ITTC, 2021. Recommended Procedures and Guidelines 7.5-04-01-01.1: Preparation, Conduct and Analysis of Speed/power Trials. URL: https://www.ittc.info/media/9874/75-04-01-011.pdf.

Jerri, A.J., 1977. The shannon sampling theorem—its various extensions and applications: a tutorial review. Proc. IEEE 65, 1565–1596.

Karagiannidis, P., Themelis, N., 2021. Data-driven modelling of ship propulsion and the effect of data pre-processing on the prediction of ship fuel consumption and speed loss. Ocean Eng. 222, 108616 https://doi.org/10.1016/j.oceaneng.2021.108616. URL: https://www.sciencedirect.com/science/article/pii/S0029801821000512.

Kim, S.-H., Roh, M.-I., Oh, M.-J., Park, S.-W., Kim, I.-I., 2020a. 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. URL: https://www.sciencedirect.com/science/article/pii/S2092678220300091.

Kim, S.-Y., Kim, M.-S., Han, K.-M., gyu Kim, J., Seo, D.-W., 2020b. Study on data analysis of on-board measurement data for ship's speed power performance. In: Ocean Engineering of International Conference on Offshore Mechanics and Arctic Engineering, 6A. https://doi.org/10.1115/OMAE2020-19153 doi:10.1115/OMAE2020-19153, v06AT06A038.

Kim, Y.-R., Jung, M., Park, J.-B., 2021. Development of a fuel consumption prediction model based on machine learning using ship in-service data. J. Mar. Sci. Eng. 9. https://www.mdpi.com/2077-1312/9/2/137. doi:10.3390/jmse9020137.

Kim, Y.-R., Esmailian, E., Steen, S., 2022. A meta-model for added resistance in waves. Ocean Eng. 266, 112749.

Kitamura, F., Ueno, M., Fujiwara, T., Sogihara, N., 2017. Estimation of above water structural parameters and wind loads on ships. Ships Offshore Struct. 12, 1100–1108.

Koboević, Ž., Bebić, D., Kurtela, Ž., 2019. New approach to monitoring hull condition of ships as objective for selecting optimal docking period. Ships Offshore Struct. 14, 95–103. https://doi.org/10.1080/17445302.2018.1481631 cited By 1.

Kristensen, H.O.H., Bingham, H., 2017. Prediction of Resistance and Propulsion Power of Ships. Technical Report, Technical University of Denmark. https://gitlab.gbar.dtu.dk/oceanwave3d/Ship-Desmo.

Lang, X., Mao, W., 2021. A practical speed loss prediction model at arbitrary wave heading for ship voyage optimization. J. Mar. Sci. Appl. 20, 410–425.

Laurie, A., Anderlini, E., Dietz, J., Thomas, G., 2021. Machine learning for shaft power prediction and analysis of fouling related performance deterioration. Ocean Eng. 234, 108886 https://doi.org/10.1016/j.oceaneng.2021.108886. URL: https://www.sciencedirect.com/science/article/pii/S0029801821003218.

Liang, Q., Tvete, H.A., Brinks, H.W., 2019. Prediction of vessel propulsion power using machine learning on AIS data, ship performance measurements and weather data. J. Phys. Conf. 1357, 12038 https://doi.org/10.1088/1742-6596/1357/1/012038 doi:10.1088/1742-6596/1357/1/012038.

Liu, S., Loh, M., Leow, W., Chen, H., Shang, B., Papanikolaou, A., 2020a. Rational processing of monitored ship voyage data for improved operation. Appl. Ocean Res. 104, 102363 https://doi.org/10.1016/j.apor.2020.102363. URL: https://www.sciencedirect.com/science/article/pii/S0141118720309226.

Liu, S., Papanikolaou, A., Feng, P., 2020b. The Semi-empirical SNNM Formula for the Approximation of Added Resistance in Waves.

Major, P., Li, G., Zhang, H., Hildre, H., 2021. Real-time Digital Twin of Research Vessel for Remote Monitoring, pp. 159–164. https://doi.org/10.7148/2021-0159.

MAN Energy Solutions, 2018. Basic principles of ship propulsion. https://www.man-es.com/docs/default-source/marine/tools/basic-principles-of-ship-propulsion_web_links.pdf?sfvrsn=12d1b862_10.

Martinsen, M.A., 2016. An Design Tool for Estimating the Added Wave Resistance of Container Ships. Master's thesis. DTU.

Minoura, M., Hanaki, T., Nanjo, T., 2021. Improvement of statistical estimation of ship performance in actual seas by normalization of data unevenness using cluster analysis. In: Okada, T., Suzuki, K., Kawamura, Y. (Eds.), Practical Design of Ships and Other Floating Structures. Springer Singapore, Singapore, pp. 878–898.

Mittendorf, M., Nielsen, U.D., Bingham, H.B., 2023. Capturing the effect of biofouling on ships by incremental machine learning. Appl. Ocean Res. 138, 103619 https://doi.org/10.1016/j.apor.2023.103619. URL: https://www.sciencedirect.com/science/article/pii/S0141118723001608.

Molland, A.F., 2011. The Maritime Engineering Reference Book: A Guide to Ship Design, Construction and Operation. Elsevier.

Munk, T., 2016. Fuel conservation through managing hull resistance. In: Motorship Propulsion Conference. Copenhagen. https://www.messe.no/ExhibitorDocuments/97726/2419/BIMCO%20Hull-Resistance.pdf.

Olmer, N., Comer, B., Roy, B., Mao, X., Rutherford, D., 2017. Greenhouse Gas Emissions from Global Shipping, 2013–2015 Detailed Methodology. International Council on Clean Transportation, Washington, DC, USA, pp. 1–38.

Olofsson, R., 2020. Unsupervised Anomaly Detection, Master's Thesis. UMEÅ University. URL: https://www.diva-portal.org/smash/get/diva2:1445794/FULLTEXT01.pdf.

Park, J., Kim, B., Jeong, S., Park, J.H., Jeong, D., Ahn, K., 2017. A comparative analysis of ship speed-power performance based on the noon reports and recorded sensor data: overcoming sensor issues. In: OCEANS 2017-Anchorage. IEEE, pp. 1–7.

Perera, L.P., Mo, B., 2018. Ship performance and navigation data compression and communication under autoencoder system architecture. J. Ocean Eng. Sci. 3, 133–143. https://doi.org/10.1016/j.joes.2018.04.002. URL: https://www.sciencedirect.com/science/article/pii/S2468013317301109.

Taskar, B., Andersen, P., 2019. Benefit of Speed Reduction for Ships in Different Weather Conditions.

Taskar, B., Andersen, P., 2021. Comparison of added resistance methods using digital twin and full-scale data. Ocean Eng. 229, 108710 https://doi.org/10.1016/j.oceaneng.2021.108710. URL: https://www.sciencedirect.com/science/article/pii/S0029801821001451.

Themelis, N., Spandonidis, C., Christopoulos, G., Giordamlis, C., 2018. A comparative study on ship performance assessment based on noon report and continuous monitoring system datasets. Proceedings 55–64.

Thomas, R., Judith, J., 2021. Hybrid dimensionality reduction for outlier detection in high dimensional data. Int. J. Emerg. Trends Eng. Res. 8 https://doi.org/10.30534/ijeter/2020/160892020.

van den Boom, H., Huisman, H., Mennen, F., 2013. New Guidelines for Speed/power Trials: Level Playing Field Established for IMO EEDI.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al., 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods 17, 261–272.

Wahl, J.M., 2019. Prediction of Fuel Consumption of a Ship in Transit Using Machine Learning. Master's thesis, NTNU. URL: https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/2622968.

Walker, M., Atkins, I., 2007. Surface Ship Hull and Propeller Fouling Managment, pp. 131–138. Cited By 3.

Wang, J., Bielicki, S., Kluwe, F., Orihara, H., Xin, G., Kume, K., Oh, S., Liu, S., Feng, P., 2021. Validation study on a new semi-empirical method for the prediction of added resistance in waves of arbitrary heading in analyzing ship speed trial results. Ocean Eng. 240, 109959 https://doi.org/10.1016/j.oceaneng.2021.109959. URL: https://www.sciencedirect.com/science/article/pii/S0029801821013020.

You, Y., Kim, J., Seo, M.-G., 2017. A feasibility study on the rpm and engine power estimation based on the combination of AIS and ECMWF database to replace the full-scale measurement. J. Soc. Naval Architects of Korea 54, 501–514. https://doi.org/10.3744/SNAK.2017.54.6.501.