Real Vessel Data Challenge: First Results, Food for Thought & Discussion

Vasileios Tsarsitalidis, ErmaTech, Athens/Greece, V.Tsarsitalidis@ermafirst.com
Dmitry Ponkratov, Siemens, London/UK, Dmitry.Ponkratov@siemens.com
Nikolaos Tsoulakos, Laskaridis Shipping, Athens/Greece, Tsoulakos@Laskaridis.com

Abstract

This paper presents the outcomes of the Real Vessel Data Challenge, a benchmarking workshop designed to compare and critically assess different methodologies applied to the same real-world dataset from a 76000 dwt bulk carrier. The challenge focused on generating speed–power baseline models, estimating performance degradation and fouling effects, reconstructing missing data, and assessing data quality and uncertainty, while remaining broadly aligned with ISO 19030. Beyond presenting comparative outcomes, the paper aims to stimulate an open technical discussion on the limits of objectivity in long-term performance analysis, the trade-off between robustness and sensitivity, and the need for clearer guidance on the application of standards to non-ideal datasets.

.

1. Introduction

The Real Vessel Data Challenge was conceived as a benchmarking exercise built around a single anonymized real-vessel dataset, with the specific aim of comparing how different analysts interpret the same operational evidence. The challenge was intentionally positioned at the intersection of ship-performance monitoring, fouling assessment, data quality control, and standard interpretation. Rather than seeking a single winning model, it was designed to expose areas of agreement and disagreement, reveal the sensitivity of results to modeling choices, and stimulate a more objective discussion around the practical use of ISO 19030 on non-ideal datasets.

The participant pool covered a diverse set of affiliations and backgrounds, including contributions from Albis Marine Performance, Chugoku Marine Paints, Nippon Kaiji Kyokai, NMRI, SRC, Lloyd's Register, MARIN, CMCC Foundation, Enamor; NTNU, SINTEF Ocean, NTU, DNV, We4Sea, NAPA, GCMD, Mærsk Mc-Kinney Møller Center for Zero Carbon Shipping, Seaspan Corporation, Foreship, RINA Group, JoRes project, and Perception (Vessel Performance Expert). The associated questionnaire responses also showed a mix of academic/research institutions, technology vendors, consultancies, startups, nonprofit bodies and mixed consortia, with declared tools ranging from Excel, MATLAB and R to Python-based regression, Bayesian, machine-learning, deep-learning and hybrid physics-informed workflows. These declarations are not used to de-anonymize the comparative discussion below, but they help explain why the submissions differed not only in numerical outputs, but also in problem framing and data-handling philosophy.

The dataset and reporting framework were accompanied by a code of conduct, confidentiality provisions, and a structured questionnaire. This provided a second layer of information beyond the technical reports: it captured declared tools, methods, organizational profiles, and baseline assumptions, which are used here to enrich the anonymized comparison. The central objective of this paper is therefore twofold: first, to compare the exported speed-power curves and the methodological choices behind them; second, to extract lessons for future benchmarking exercises and for the continued interpretation and possible refinement of ISO 19030.

2. Challenge Scope, Submission Set and Harmonization Procedure

2.1. Submission set and quantitative summary

The full participant set comprised 13 teams and 14 documented entries, because one team submitted two distinct machine-learning models (there was the option for multiple entries per team, but only one exercised this). The quantitative comparison presented here was limited to the exported speed-power

curves contained in the submitted CSV files. Since the exports were not fully standardized, an organizer-side harmonization step was required before any common plotting was possible.

Table I: Submission and harmonization summary used in the quantitative comparison

Participating teams	13
Documented entries discussed in the paper	14
Entries represented by exported speed-power curves	14
CSV files processed	49
Retained speed-power points	3624

2.2. Harmonization of submitted curve data

The common-curve comparison was based on all exported files containing explicit speed-power relationships, whether submitted as single consolidated datasets or as one file per draft, condition, or temporal variant. Organizer-side harmonization normalized the speed coordinate to knots and the power coordinate to kW, preserved participant labels, and introduced a condition-family layer used only for comparison plots.

Condition naming was one of the first issues revealed by the exercise. Some participants used explicit clean/current labels; others used fouled, baseline, newbuilt, December 2024, or month-specific labels. The resulting families were therefore grouped as clean-exact, current-like, baseline-like, monthly-variant, and unlabeled special cases. This was not done to erase semantic differences, but to make those differences visible in a manageable visual structure.

For the mean comparison line in each common plot, pointwise outlier removal was first applied and a smooth exploratory mean curve was then constructed. The resulting mean curves are therefore useful anchors for discussion, but not authoritative physical references. In addition, descriptive deviation scores relative to the smoothed family mean were calculated to test whether simple spread metrics can help summarize agreement and disagreement. These scores are used here descriptively only and are not interpreted as rankings.

2.3. Updated scope

In order to produce a commonly acceptable ranking, some “ground truth” would have to be established. Such a ground truth, could be a high fidelity CFD baseline model, but this has not been agreed upon with the participants and the scoring of the challenge (and the consequential declaration of the winner) will be concluded after open deliberation with all the participants. Thus, this paper will be limited to the anonymized comparison of the submissions among themselves, and the “competitive” part of the challenge will be shown in the presentation, after proper discussion with all participants.

3. Comparative Methodological Review

A key outcome of the challenge is that the entries differed not only in algorithm, but in problem definition. Some entries approached the task as a largely ISO 19030-compliant speed-power and degradation assessment. Others treated it first as a sensor-diagnostics problem. A third group framed it as a physics-based baseline reconstruction exercise, while another group treated it primarily as a statistical or machine-learning estimation task. The questionnaire responses reinforce this distinction by showing materially different declared tools and modeling traditions behind the entries. This distinction is essential because it explains why disagreement between submissions should not immediately be read as disagreement between models alone. In several cases, the participants were solving subtly different inverse problems with the same data. The condensed problem-definition table therefore includes both the report-based analytical interpretation and a short questionnaire-derived profile for each anonymized entry.

Table II: Condensed problem-definition summary for all entries discussed in the paper

	Questionnaire profile	**Effective question being solved**	**Main trust anchor**	**Main limiting factor**	**ISO-related implication**
Entry 1	Consult./eng.; Matlab; regressions + hybrid	Can ISO-like processing recover comparable outputs?	Standards workflow	Weak coverage after aggressive filtering	Needs clearer guidance for non-ideal datasets
Entry 2	Acad./research; Python; regressions + physics + Bayesian	Can monthly shifts be detected robustly?	Bayesian consistency bands	Attribution ambiguity between fouling and instrumentation	Detected shift should be distinguished from confirmed fouling
Entry 3	Tech vendor; Excel/SQL; ML + physics + hybrid	Can normalized trends support blind prediction?	Availability checks + hybrid model	Incomplete blind dataset	Define handling of blind/incomplete continuation periods
Entry 4	Acad./research + tech vendor; Excel/OCTARVIA; regressions + physics	How do actual-sea curves compare to clean baselines?	Towing-tank + sea-trial baseline	Limited sensor critique in report	Bridge still-water and actual-sea benchmarking more explicitly
Entry 5	Acad./research; Python; regressions	Can HF data be statistically condensed?	Hourly aggregation + OLS transparency	Smoothing may suppress fouling	Clarify acceptable temporal aggregation
Entry 6	Tech vendor; Python; DL + physics + hybrid	Can sensor diagnostics precede fouling inference?	Shaft-power/RPM envelopes + SFOC	Recalibration + state imbalance	Add explicit sensor-health stage
Entry 7	Non-profit; Python; regressions + physics + hybrid	Can corrected calm-water curves reveal fouling?	Cross-sensor engineering checks	Post-July consistency failure	Require torque/power/rpm cross-validation
Entry 8	Tech vendor; R; regressions + hybrid	Can gradual fouling be modeled as time drift?	Physics baseline + regression	Limited case-specific audit	Hybrid formulations may deserve more explicit recognition
Entry 9	Small team; Python; PINN + physics + hybrid	Can physical feasibility improve extrapolation?	Physical decomposition in AI	Too little detail for validation	Useful philosophy, limited evidence here
Entry 10	Acad./research consortium; Python/Excel; regressions + physics + hybrid	What remains possible under weak metadata?	Engineering standards + checks	Missing metadata and sensor trust	Specify minimum metadata/sensor verification requirements
Entry 11	Tech vendor/startup; Python/Excel; regressions + physics	Can fuel benchmarking work without full HF engine detail?	Digital twin + noon fuel	Noon timing uncertainty	Consider a fuel-centric pathway
Entry 12	Acad./research + consultancy; Python; regressions + ML	Can fouling proxies be learned from HF data?	Large preprocessed HF set	Missing GPS/draft/context	State more clearly how missing context limits objectivity
Entry	Acad./research +	Same with	Proxy + SHAP	Location blindness	Allow hybrid data-

13	consultancy; Python; regressions + ML	nonlinear flexibility	interpretability	+ collinearity	driven methods with stricter assumption reporting
Entry 14	Tech vendor/consultancy; Python/Tableau; regressions + physics + hybrid	Can heavy corrective preprocessing recover usable baselines?	Corrected variables + model-test guidance	Raw variables too compromised	Better address corrective preprocessing of incomplete inputs

4. Quantitative Comparison of Exported Curves

The main task of the challenge was to provide baseline curves for the clean and current condition of the vessel, given the real time data of a whole year along with the sea trials curves and main particulars of the vessel.

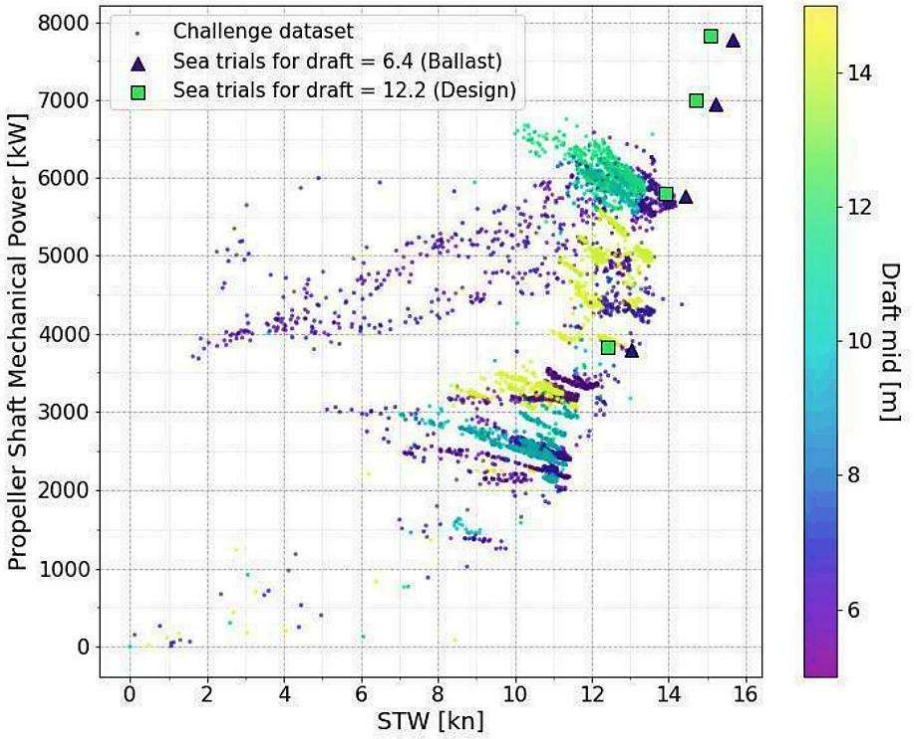

Fig.1: All speed vs power data of the dataset given to participants (shown to indicate the complexity)

The harmonized curve comparison reinforces the methodological findings. Once the exported curves are grouped by practical condition families, the plots show both where the entries converge and where they diverge due to semantics, baseline philosophy, or possible data-treatment differences. The intention of the figures is therefore diagnostic rather than competitive.

The common plots were built to reveal issues rather than suppress them. Each figure overlays anonymized submissions, grouped by draft where possible, and adds an exploratory smoothed family mean after pointwise outlier removal (dashed line). This allows families with reasonable overlap to be visually compared while still preserving outliers and special branches that may themselves be informative.

Table III: Exploratory quantitative summary of the harmonized curve families. The deviation metrics are descriptive only and are not used as rankings.

	Entries	Files	Points	Drafts represented	Median MAPE to mean	Median RMSE to mean
clean t	8	38	1,679	Ballast, Heavy ballast, 10 m, 11 m, Design, Scantling	5.5%	293 kW
current	7	8	802	Ballast, Heavy ballast, 10 m, 11 m, Design, Scantling	8.8%	470 kW
monthly variant	2	2	245	Ballast, Scantling	4.7%	154 kW

4.1. Speed VS Power Curves

Figs.2 to 6 show the collected curves for the clean and current condition (columns) and for the different drafts (lines). The clean-condition family shows the best overlap in the better-populated draft bands, but even there the spread is not negligible. This indicates that the challenge did not produce a single clean-baseline consensus. Some entries were close in both curvature and absolute power level, while others remained systematically higher or lower, reflecting different assumptions on what constituted clean reference behavior. The current condition family is more heterogeneous still, because it pools labels such as current, fouled, and December 2024. The common plot makes clear that current was not defined uniformly across participants. In several drafts the spread is wider than in the clean family, which is consistent with the report-based finding that sensor-health interpretation and baseline semantics were major differentiators.

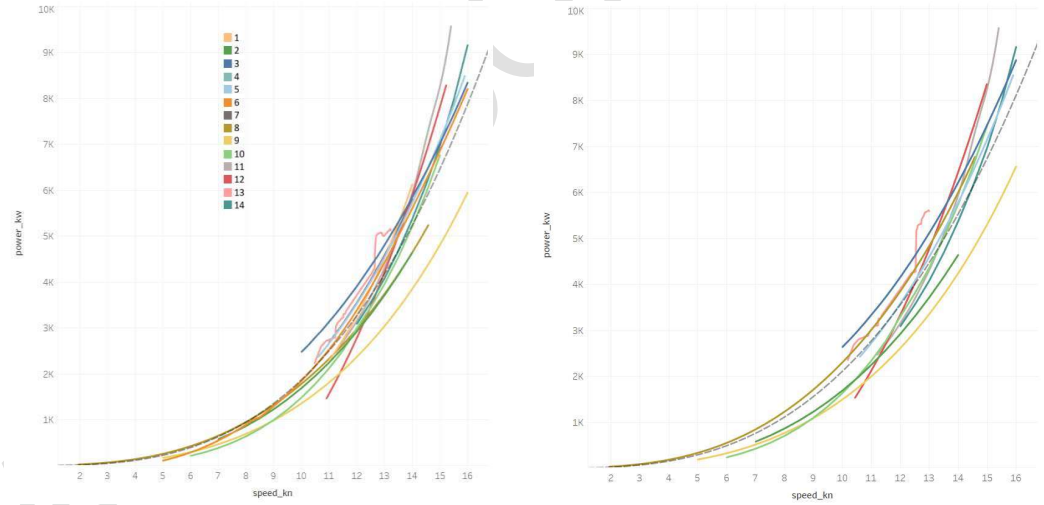

Fig.2: Curves for clean hull (left) and current condition (right) at ballast draft

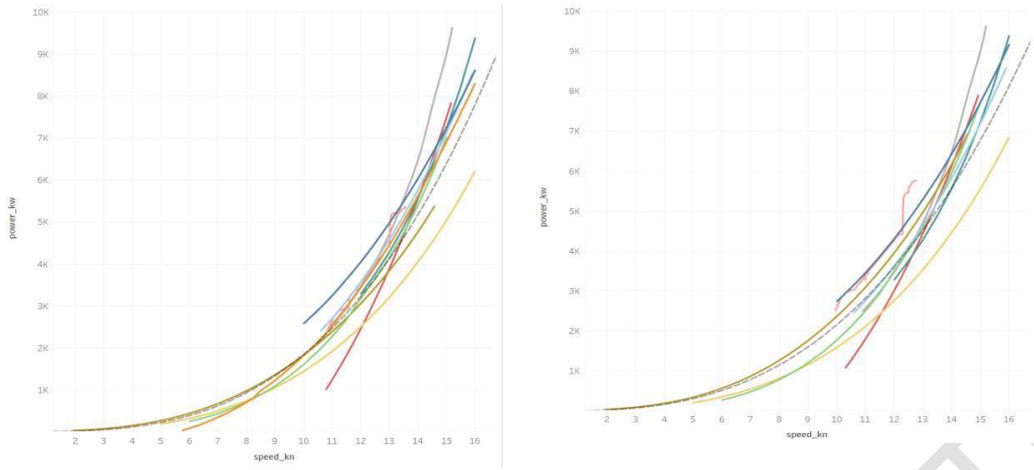

Fig.3: As Fig.2, at heavy ballast draft

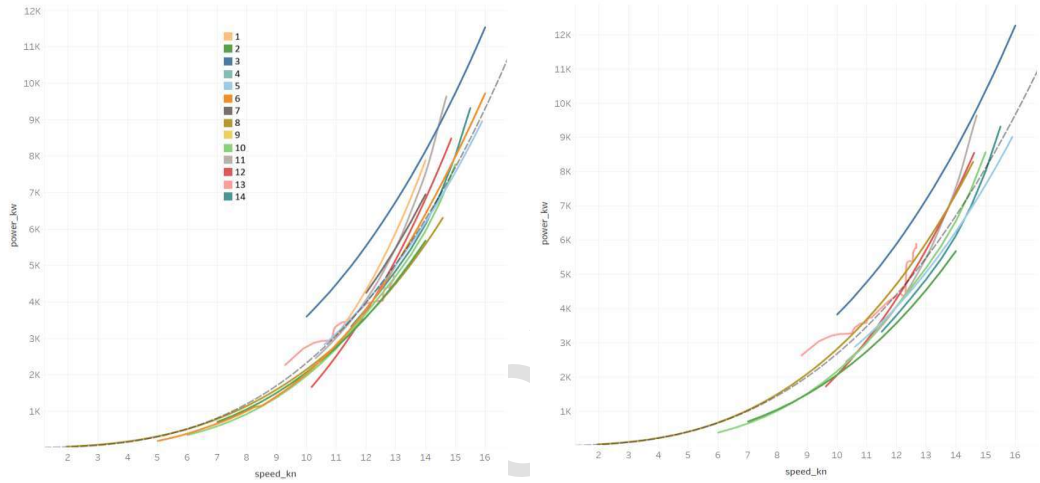

Fig.4: As Fig.2, at design draft

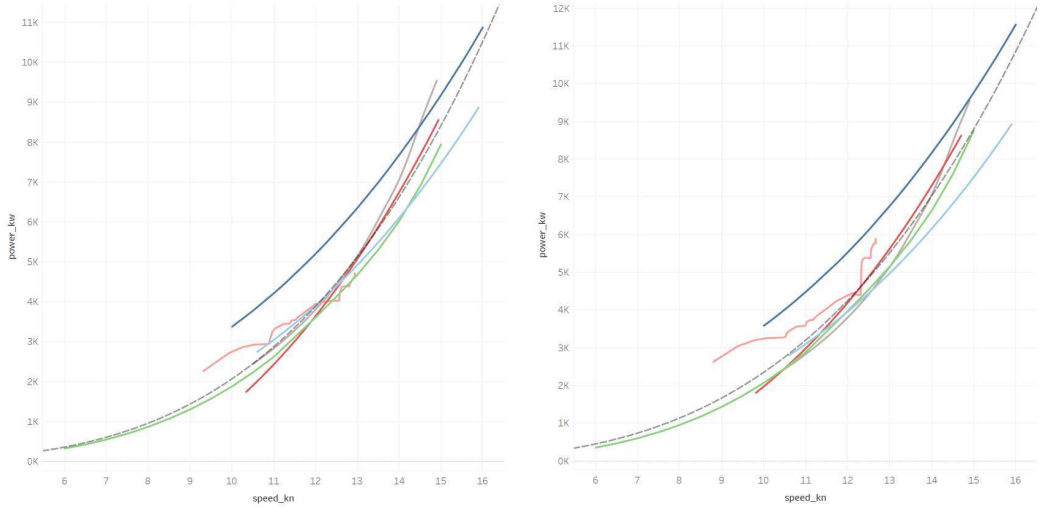

Fig.5: As Fig.2, at 11 m draft

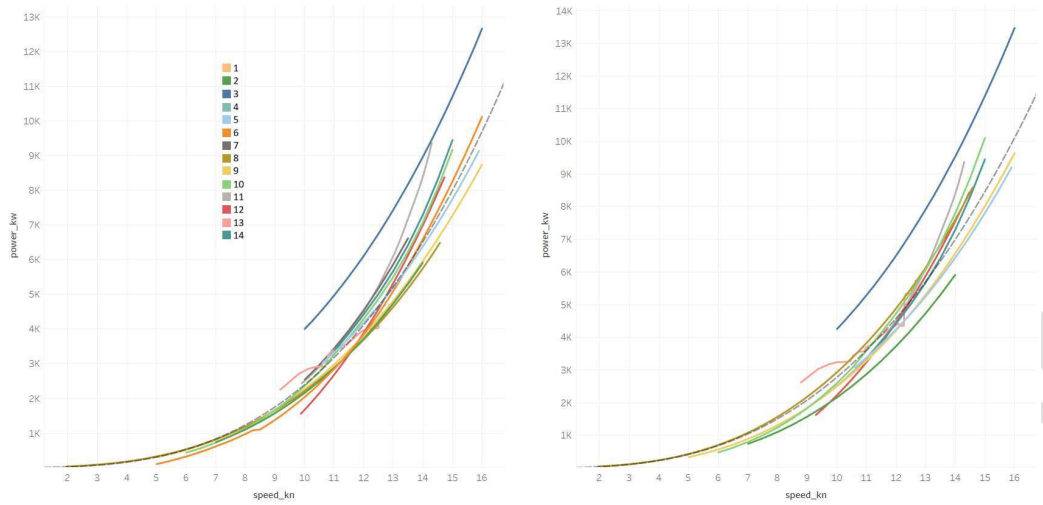

Fig.6: As Fig.2, at scantling draft

5. Discussion

The curve comparison and the report review point in the same direction: the largest differences between entries are not simply algorithmic. They arise much earlier, in how participants define the problem, choose or reconstruct baselines, decide which channels are trustworthy, and determine whether a shift in the data is operational, hydrodynamic, or instrumental.

A second major result is that data trustworthiness often prevails over model sophistication. Several entries independently pointed to structural breaks in torque- or propulsion-related signals around mid-2024, although they framed these breaks differently: recalibration, maintenance, or related sensor inconsistency. This convergence is important because it suggests that part of the challenge became an exercise in diagnosing instrumentation consistency before interpreting fouling trends.

The quantitative curve plots also show that direct averaging can be misleading when baseline semantics differ. The exploratory family mean is useful as a visual anchor, but it should not be interpreted as a physical reference line whenever the family itself mixes different clean or current philosophies. In that sense, the mean curve is helpful precisely because it highlights where averaging ceases to be conceptually safe. These observations also explain why the challenge should not be interpreted as a ranking exercise. The descriptive deviation scores are useful for testing spread metrics, but they cannot decide whether a curve is good or bad without reference to a common baseline philosophy and a common view of data validity. The challenge was therefore more valuable as a structured comparison of problem framings than as a competition for a single best line.

6. Implications for ISO 19030

The reports, the exported curve files, and the harmonized comparison together suggest several practical improvements or clarifications relevant to ISO 19030 and to the broader practice of long-term performance assessment on operational data. The points below combine the report review with the issues surfaced by the quantitative curve harmonization.

1. Separate performance-change detection from fouling attribution. A detected shift should not automatically be labeled fouling unless maintenance and instrumentation effects are excluded.

2. Add an explicit sensor-health stage. Torque-meter recalibration, unit inconsistencies, and cross-sensor propulsion checks should appear before fouling assessment, not only as informal pre-checks.
3. Provide clearer guidance for non-ideal datasets. Missing draft, missing GPS, uncertain noon timing, corrupted propulsion channels, and incomplete blind periods were all material in this challenge.
4. Clarify acceptable temporal aggregation. The challenge included 10-minute, 15-minute, 20-minute, hourly, and daily approaches, each with different implications for noise rejection and trend meaning.
5. Standardize weather and wave-treatment practice more clearly. Weather-source choice, wave filtering, and wave correction materially influenced the results.
6. Tighten minimum reporting requirements. Baseline definition, resampling logic, weather source, sensor corrections, uncertainty method, and synthetic inputs should always be disclosed.
7. Clarify acceptable clean-baseline concepts. Sea-trial, towing-tank, digital-twin, statistical, and post-cleaning baselines are not interchangeable and should not be treated as such.
8. Develop a common operational uncertainty language. The reports quantified very different uncertainties using very different tools and assumptions.
9. Be cautious about hull/propeller separation. Simultaneous cleaning and correlated proxies can make this distinction unreliable.
10. Consider a fuel-centric benchmarking pathway. Some operationally meaningful methods are not naturally shaft-power-centric and should not be excluded from comparison frameworks.

7. Conclusions

The first Real Vessel Data Challenge succeeded in demonstrating that common access to a single real-vessel dataset does not guarantee common interpretation. The participant reports, the questionnaire responses, and the harmonized curve comparison together show that variability arises from baseline philosophy, data-quality diagnosis, temporal treatment, environmental correction, and the willingness to question the instrumentation before trusting the trend. The report review and the curve comparison both indicate that the most important source of disagreement is often upstream of the final model. Sensor trust, missing context, incomplete metadata, and semantic differences in what counts as clean or current can dominate the apparent outcome. This is not a weakness of the challenge; it is one of its most useful results.

Future iterations of the challenge should preserve this openness while tightening the numerical export schema and minimum reporting requirements. More rigorous harmonization rules, richer metadata, and clearer distinction between change detection and attribution would make future comparisons sharper without suppressing the methodological diversity that made this first exercise valuable.

Acknowledgement

We gratefully acknowledge the participants of the Real Vessel Data Challenge, Laskaridis Shipping, the donor side for enabling the anonymized release of the dataset, and the broader HullPIC community for providing the forum in which these results can be discussed openly.

References

DNV (2023), *Recommended practice for verification and use of ship performance models*, DNV-RP-675, DNV, Hovik

ISO (2016), *ISO 19030 - Measurement of changes in hull and propeller performance*, Int. Standard Org., Geneva

ISO (2025), *ISO 15016 - Guidelines for speed and power trials*, Int. Standard Org., Geneva

Appendix A: Master comparison of submitted entries

Table A1: Master comparison of all entries in anonymized form, enriched with questionnaire-derived profile information

	Questionnaire profile	**Problem framing**	**Method class**	**Main data-issue diagnosis**	**Data-quality stance**	**Practical takeaway**
Entry 1	Consult./eng.; Matlab; regressions + hybrid	ISO-style recovery of clean/current/fouling outputs	Standards-led engineering	No single dominant instrumentation fault isolated; strict filtering/correction central	Formal, staged, standards-driven	ISO usable, but practical deviations were needed
Entry 2	Acad./research; Python; regressions + physics + Bayesian	Monthly performance-shift inference with uncertainty	Bayesian operational modelling	July shift may reflect recalibration or maintenance, not only fouling	Cautious, uncertainty-aware	Shift detected, cause not uniquely attributable
Entry 3	Tech vendor; Excel/SQL; ML + physics + hybrid	Reference-normalized fouling diagnosis and blind prediction	Hybrid empirical + ML	Blind data incomplete; draft and heading assumptions required	Pragmatic, availability-focused	October peak underperformance; later partial recovery
Entry 4	Acad./research + tech vendor; Excel/OCTARVIA; regressions + physics	Voyage-wise comparison to clean physical baselines	Physics/software operational-sea	Main challenge is voyage grouping and fit quality	Objective-fit, voyage-based	Actual-sea comparison feasible when clean baseline exists
Entry 5	Acad./research; Python; regressions	Statistical recovery of vessel-performance curve	Interpretable OLS	HF data too sparse/noisy for direct use	Conservative, statistical	Hourly smoothing robust, but fouling term weak
Entry 6	Tech vendor; Python; DL + physics + hybrid	Sensor diagnosis first, fouling second	Grey-box hybrid	Torquemeter inconsistency/recalibration after July	Highly diagnostic	Sensor-health correction is essential
Entry 7	Non-profit; Python; regressions + physics + hybrid	Calm-water curve after environmental correction	Engineering correction + regression	20–25% KQ step change after July	Strong engineering validation	Uncorrected split-year comparison not meaningful
Entry 8	Tech vendor; R; regressions + hybrid	Fouling as time drift on physics baseline	Physics + empirical hybrid	No specific anomaly emphasized	Structural/model-based	Useful hybrid framework; more method note than audit
Entry 9	Small team; Python; PINN + physics + hybrid	Physically feasible ML extrapolation	Hybrid hydrodynamics + ML	Too little case-specific detail	Method-focused	Insufficient detail for full comparison
Entry 10	Acad./research consortium;	Engineering procedure	Physics/engineering	Possible turbocharger or	Skeptical, realism-	Benchmark exposed

	Python/Excel; regressions + physics + hybrid	under incomplete metadata	procedures	mass-flow-meter issue; missing metadata	driven	sensor-trust and metadata limits
Entry 11	Tech vendor/startup; Python/Excel; regressions + physics	Fuel-performance benchmarking vs digital twin	Digital twin / fuel-centric	Noon timing uncertain; blind period narrow	Aggregation - and availability-centered	Operationally relevant but only partly comparable to power-centric methods
Entry 12	Acad./research + consultancy; Python; regressions + ML	Learn performance using fouling roughness proxy	ML with engineered proxy	Missing draft/GPS; corrupted propulsion channels	Explicit issue-register style	Promising, but context limitations remain strong
Entry 13	Acad./research + consultancy; Python; regressions + ML	Same proxy problem with nonlinear learning	ML / nonlinear	Same limits, plus hull/prop roughness collinearity	Issue-register + SHAP	Interpretable ML possible, but context still limiting
Entry 14	Tech vendor/consultancy; Python/Tableau; regressions + physics + hybrid	Parallel data-driven and model-test baselines	Regression + model-test hybrid	Major corrections needed for STW, power, torque, and swell descriptors	Correction-led	Preprocessing burden is itself a key result

Appendix B: Problem-definition summary

Table B1: Problem-definition table used in the main discussion

	Questionnaire profile	**Effective question being solved**	**Main trust anchor**	**Main limiting factor**	**ISO-related implication**
Entry 1	Consult./eng.; Matlab; regressions + hybrid	Can ISO-like processing recover comparable outputs?	Standards workflow	Weak coverage after aggressive filtering	Needs clearer guidance for non-ideal datasets
Entry 2	Acad./research; Python; regressions + physics + Bayesian	Can monthly shifts be detected robustly?	Bayesian consistency bands	Attribution ambiguity between fouling and instrumentation	Detected shift should be distinguished from confirmed fouling
Entry 3	Tech vendor; Excel/SQL; ML + physics + hybrid	Can normalized trends support blind prediction?	Availability checks + hybrid model	Incomplete blind dataset	Define handling of blind/incomplete continuation periods
Entry 4	Acad./research + tech vendor; Excel/OCTARVIA; regressions + physics	How do actual-sea curves compare to clean baselines?	Towing-tank + sea-trial baseline	Limited sensor critique in report	Bridge still-water and actual-sea benchmarking more explicitly
Entry 5	Acad./research; Python; regressions	Can HF data be statistically condensed?	Hourly aggregation + OLS transparency	Smoothing may suppress fouling	Clarify acceptable temporal aggregation
Entry 6	Tech vendor; Python; DL + physics + hybrid	Can sensor diagnostics precede fouling inference?	Shaft-power/RPM envelopes + SFOC	Recalibration + state imbalance	Add explicit sensor-health stage
Entry 7	Non-profit; Python; regressions + physics + hybrid	Can corrected calm-water curves reveal fouling?	Cross-sensor engineering checks	Post-July consistency failure	Require torque/power/rpm cross-validation
Entry 8	Tech vendor; R; regressions + hybrid	Can gradual fouling be modeled as time drift?	Physics baseline + regression	Limited case-specific audit	Hybrid formulations may deserve more explicit recognition
Entry 9	Small team; Python; PINN + physics + hybrid	Can physical feasibility improve extrapolation?	Physical decomposition in AI	Too little detail for validation	Useful philosophy, limited evidence here
Entry 10	Acad./research consortium; Python/Excel; regressions + physics + hybrid	What remains possible under weak metadata?	Engineering standards + checks	Missing metadata and sensor trust	Specify minimum metadata/sensor verification requirements
Entry 11	Tech vendor/startup; Python/Excel; regressions + physics	Can fuel benchmarking work without full HF engine detail?	Digital twin + noon fuel	Noon timing uncertainty	Consider a fuel-centric pathway
Entry 12	Acad./research +	Can fouling	Large	Missing	State more

	consultancy; Python; regressions + ML	proxies be learned from HF data?	preprocessed HF set	GPS/draft/context	clearly how missing context limits objectivity
Entry 13	Acad./research + consultancy; Python; regressions + ML	Same with nonlinear flexibility	Proxy + SHAP interpretability	Location blindness + collinearity	Allow hybrid data-driven methods with stricter assumption reporting
Entry 14	Tech vendor/consultancy; Python/Tableau; regressions + physics + hybrid	Can heavy corrective preprocessing recover usable baselines?	Corrected variables + model-test guidance	Raw variables too compromised	Better address corrective preprocessing of incomplete inputs

Appendix C: ISO 19030 improvement points emerging from the challenge

1. Separate performance-change detection from fouling attribution. A detected shift should not automatically be labeled fouling unless maintenance and instrumentation effects are excluded.
2. Add an explicit sensor-health stage. Torque-meter recalibration, unit inconsistencies, and cross-sensor propulsion checks should appear before fouling assessment, not only as informal pre-checks.
3. Provide clearer guidance for non-ideal datasets. Missing draft, missing GPS, uncertain noon timing, corrupted propulsion channels, and incomplete blind periods were all material in this challenge.
4. Clarify acceptable temporal aggregation. The challenge included 10-minute, 15-minute, 20-minute, hourly, and daily approaches, each with different implications for noise rejection and trend meaning.
5. Standardize weather and wave-treatment practice more clearly. Weather-source choice, wave filtering, and wave correction materially influenced the results.
6. Tighten minimum reporting requirements. Baseline definition, resampling logic, weather source, sensor corrections, uncertainty method, and synthetic inputs should always be disclosed.
7. Clarify acceptable clean-baseline concepts. Sea-trial, towing-tank, digital-twin, statistical, and post-cleaning baselines are not interchangeable and should not be treated as such.
8. Develop a common operational uncertainty language. The reports quantified very different uncertainties using very different tools and assumptions.
9. Be cautious about hull/propeller separation. Simultaneous cleaning and correlated proxies can make this distinction unreliable.
10. Consider a fuel-centric benchmarking pathway. Some operationally meaningful methods are not naturally shaft-power-centric and should not be excluded from comparison frameworks.