Benchmarking models for ocean wind speed estimation based on passive acoustic monitoring

1st Anatole Gros-Martial
Centre d'Etudes Biologiques de Chizé
CNRS
Villiers-en-Bois, France
grosmartial.anatole@gmail.com

2nd Louise Delaigue
Laboratoire d'Océanographie de Villefranche
Sorbonne Université, CNRS
Villefranche-sur-mer, France
louise.delaigue@imev-mer.fr

3rd Pierre Cauchy
Institut des sciences de la mer (ISMER)
Université du Québec à Rimouski (UQAR)
Rimouski, Canada
pierre_cauchy@uqar.ca

4th Sara Pensieri
Institute for the Study of Anthropic Impact and Sustainability in the Marine Environment (IAS)
Consiglio Nazionale delle Ricerche (CNR)
Genoa, Italy
sara.pensieri@cnr.it

5th Roberto Bozzano
Institute for the Study of Anthropic Impact and Sustainability in the Marine Environment (IAS)
Consiglio Nazionale delle Ricerche (CNR)
Genoa, Italy
roberto.bozzano@cnr.it

6th Julien Bonnel
Applied Ocean Physics and Engineering Department
Woods Hole Oceanographic Institution
Woods Hole, MA 02543, USA
jbonnel@whoi.edu

7th Edouard Leymarie
Laboratoire d'Océanographie de Villefranche
Sorbonne Université, CNRS
Villefranche-sur-mer, France
edouard.leymarie@imev-mer.fr

8th Mark Baumgartner
Biology Department
Woods Hole Oceanographic Institution
Woods Hole, MA 2543, USA
mbaumgartner@whoi.edu

9th Christophe Guinet
Centre d'Etudes Biologiques de Chizé
CNRS
Villiers-en-Bois, France
christophe.guinet@cebc.cnrs.fr

10th Sara Bazin
Geo-Ocean, IUEM
Univ Brest, CNRS, Ifremer
Plouzané, France
sara.bazin@univ-brest.fr

11th Dorian Cazau
ENSTA
Lab-STICC, UMR CNRS 6285
Brest, France
dorian.cazau@ensta.fr

***Abstract*—Passive acoustic monitoring is a promising tool for long-term ocean observations, offering a unique means to capture physical and biological processes. This study explores its potential as a source of fine temporal scale in-situ wind speed product by assembling a unique corpus of acoustic datasets co-located with or near in-situ weather stations. This study offers two key contributions : *i)* setting up a benchmarking framework for the development and evaluation of models in acoustic meteorology, and *ii)* applying this framework to assess the performance of various models, comparing parameters from the literature with those trained on datasets from this study's corpus. Regarding the latter point, results show that most untrained models fail to generalize due to the intrinsic variability of soundscape in different basins and environmental conditions, as well as calibration inaccuracies. However, all models can achieve satisfactory performance on specific datasets after training. Incorporating diverse observational sources, such as gliders and BGC-Argo floats, could enhance model robustness, and improved acoustic-based estimates will help refine satellite-derived wind products and numerical weather predictions, ultimately advancing global wind field modeling and air-sea interaction research.**

***Index Terms*—Acoustic meteorology, wind generated ocean noise, wind speed estimation, passive acoustic monitoring**

The research leading to these results has received part of the funding from the European Union's Horizon research and innovation program under grant agreement 101008724 (MINKE project) and 101188028 (TRICUSO project), from the JPI Ocean Initiative Underwater noise in the marine environment and the Italian Ministry of University and Research under the PURE WIND project.

I. Introduction

Passive acoustic monitoring (PAM) is a diverse and promising data source to detect and analyze ocean soundscapes. Its applications range from sounds produced by fish and marine mammals to ambient noise from physical oceanographic pro-

cesses. Despite its important range of applications, PAM is still rarely used for operational ocean observations. The recent proposal of an ocean sound-based Essential Ocean Variable [1] within the Global Ocean Observing System (GOOS) initiative sets the stage for a more comprehensive and operational incorporation of PAM into long-term international observation programs, such as Euro-Argo and EMSO. With this ambition in mind, the present research focuses on benchmarking current analytical methods to use PAM as a source of fine temporal scale in-situ surface wind speed products.

This new source of in-situ wind speed measurement would fill an important gap in ocean sampling. For example, the World Meteorological Organization, in support of Earth Observation applications, emphasizes that key meteorological and oceanographic fields, such as Numerical Weather Prediction (NWP) and Met-Ocean Forecasts and Services, rely on precise and high-resolution observations of met-ocean parameters. Among the observational requirements, there is a dire need for frequent real-time observations with better spatial-resolution and temporal sampling less than three hours for global NWP. Nowadays, conventional in-situ weather measurement methods, such as ship-based stations (including academic, commercial, and volunteer vessels) and sensor-equipped drifting or moored buoys, provide spatially heterogeneous and often sparse oceanic coverage. This limitation underscores the need for alternative data sources to enhance the resolution and frequency of ocean observations. Beyond its meteorological significance, wind speed plays a critical role in air-sea gas exchanges, particularly in the flux of uptake CO_2 between the atmosphere and the ocean. Wind-driven turbulence and wave-breaking processes enhance the exchange of gases across the air-sea interface, influencing global carbon cycles and ocean biogeochemistry [2]–[4]. Consequently, improving in-situ wind speed measurements through alternative methods could contribute not only to weather prediction but also to a better understanding of oceanic uptake CO_2 climate regulation [5], [6].

In 1948, Knudsen's pioneering research [7] demonstrated a direct correlation between Beaufort wind speed and the overall underwater noise level. This has led to the development of acoustic meteorology models, hereby referred to as Weather Observation Through Ambient Noise (WOTAN) [8] techniques, where wind speed, rainfall [9]–[12] and even snow [13] can be estimated from ambient noise. The relationship described by Knudsen is caused by the generation of wind-induced breaking waves and whitecaps, which produce a layer of oscillating bubbles at the ocean's surface. These bubbles act as the primary and dominant source of sound, and their recordings can be used to infer wind speed. As wind speed increases, the prevalence and intensity of breaking waves and whitecaps grow, leading to a thicker layer of bubbles and a corresponding rise in underwater ambient noise [14]–[16]. Since then, multiple studies have developed models to estimate surface wind speed from underwater acoustic datasets, paving the way towards operational weather systems.

WOTAN platforms have several interesting advantages over traditional in-situ weather systems. The first is their immersive geometry, which makes them less sensitive than weather buoys to extreme weather conditions. For example, biologged southern elephant seals have been envisioned as potential WOTAN platforms below the 60° parallel [21]. Furthermore, embedding WOTAN models in similar multi-sensor platforms can provide weather conditions precisely co-located with other oceanographic measurements (e.g. studying the dependence of biogeochemical parameters on sea state using autonomous platforms such as BGC-Argo floats).

However, WOTAN estimations are far from being used systematically and officially in global observation programs like GOOS, because they still lack the key robustness and generalization capabilities of operational measurement systems to operate seamlessly across different marine environments. One of the main factors responsible for this is the methodology used to develop and evaluate WOTAN models. Indeed, many studies have either developed and evaluated models within a single experimental observation setup, or against coarse-resolution time series as provided by data assimilation models (*e.g.* ECMWF - one hour time resolution and 0.25° spatial resolution), but with the consequence of developing overfitted or smoothed models without adding any real value for marine meteorology. Thus, the generalization performance of models across different sites, time periods and observation platforms still needs to be assessed through more systematic and large-scale benchmarking works. Such benchmarks need precise ground truth data to be properly developed and evaluated. Ideally, reference measurements come from in-situ systems so as to make the most of WOTAN models in terms of estimation accuracy and scale of temporal resolution. Furthermore, these reference datasets should be diverse in terms of recording devices, soundscapes, and environmental conditions. This diversity enforces models to develop generalization capabilities in space and time to unseen data, which is crucial considering the multiplicity of variability factors from one dataset to another, due for example to the high site-dependency of the ambient noise - pressure level relation [22] and the presence of multiple other sound sources in the ocean that skew predictions [14], [23]. Such requirements make the task of gathering a corpus of datasets very cumbersome and difficult.

In this context, the present large-scale and comprehensive benchmarking study aims to better understand and evaluate the capacity of current WOTAN analytical models to estimate wind speeds above sea surface. To our knowledge, this study has gathered the largest corpus of PAM datasets with nearby in-situ wind speed ground truth dedicated to acoustic meteorology. All datasets were selected to include at best a large diversity of sites, time periods, recording platforms, etc. Five different types of analytical models found in literature were selected and implemented [8], [17]–[20]. This benchmark provides a standardized framework for evaluating and comparing these different methodologies, fostering advancements in wind estimation from underwater acoustic data. Three numerical experiments are introduced to compare model performance, through two evaluation routines and two metrics - applying

models with parameters from the literature, adding an ambient noise offset, and training all parameters of the models on the dataset corpus. Consideration is given to inherent differences in the deployment depth and environment as well as the instrument's calibration.

II. Material and Methods

A. Datasets

Several acoustic datasets (metadata available in Tab II) from different locations around the globe and in different environmental conditions were gathered, all associated with a weather station providing in-situ wind speed ground truth. Co-location distances range from 0 to 30 km. Anemometers are mounted on each weather station at varying heights, giving U10 equivalent measurements.

TABLE I: Beaufort wind classes

Beaufort wind class	Minimum wind speed (m/s)	Maximum wind speed (m/s)
B0	0.0	0.3
B1	0.3	1.6
B2	1.6	3.4
B3	3.4	5.5
B4	5.5	8.0
B5	8.0	10.8
B6	10.8	13.9
B7	13.9	17.2
B8	17.2	20.8
B9	20.8	24.5
B10	24.5	28.5
B11	28.5	32.7

From the acoustic datasets, two are fixed hydrophones from the DCLDE 2015 workshop (SOCAL_E, CINMS_B)[1], two are moored platforms (W1M3A [2], BOUSSOLE [3]), with W1M3A being co-located with an anemometer, and one is a moving platform (glider WHOI [4]). They were all calibrated according to the instrument's sensitivity, gain and peak-to-peak ADC value and absolute Power Spectral Densities (PSD) in $dB\ re\ 1\ \mu Pa^2/Hz$ were computed (see Figure 1).

Datasets, except W1M3A which was already processed, were formatted and processed using the open source python library *OSEkit*[5]. They were resampled at 32000 Hz and cut into 60 s chunks. From these audio recordings, Fast Fourier Transforms were computed with an $NFFT$ of 1024 and no overlap. [8] trained their models using the intensity level within a 1 Hz bandwidth centered at a frequency f, as it efficiently characterizes ambient noise which is broadband and with an important dynamic range. Following this approach, the final noise level used for surface wind speed estimation is the PSD centered at 8000 Hz with a frequency bandwidth of 5 Hz.

B. Evaluation protocols

The following evaluation protocols are applied throughout the study. The metric used is the mean absolute error (MAE) because it provides a direct measure of average prediction error, ensuring interpretation in the same units as the target variable. Compared to squared error metrics, MAE is less sensitive to outliers, making it a robust choice for evaluating model performance under diverse oceanic conditions. For each protocol, the predictions are converted to Beaufort wind force classes (Tab. I), and an accuracy score and balanced accuracy score (BAS) are computed.

1) Stratified K-fold Cross-Validation (S-KCV): In this evaluation protocol, the dataset is partitioned into training and testing sets using a stratified K-fold cross-validation algorithm based on the distribution of Beaufort wind force classes. The data are divided into five splits, with four splits used for training and one split reserved for testing in each iteration. This ensures that the distribution of wind force classes is preserved across all splits. After training the model on each split, error metrics from the individual splits are averaged to yield the final model metrics.

2) Timeseries K-fold Cross-Validation (T-KCV): The timeseries K-fold cross-validation protocol evaluates model performance over temporal splits of the dataset. The data are split into five sequential temporal splits, where each split contains 20% of the total dataset. For each evaluation, the training set comprises of four splits making up 80% of the temporal data, while the testing set corresponds to the last 20%. This iterative process ensures that every data point is eventually included in the testing set. By preserving the temporal structure of the data, this protocol captures the chronological dependencies within the dataset. After training the model in each split, error metrics from the individual splits are averaged to yield the final model metrics.

TABLE II: Dataset corpus of acoustic datasets and their distance to the wind speed ground truth. Bathymetry is obtained approximately using GEBCO2023 data at given hydrophone deployment position.

	Latitude	Longitude	Distance to ground truth (km)	Depth (m)	Bathymetry (m)
BOUSSOLE	43.4	7.9	6.08	15	2388
SOCAL_E	32.7	-119.5	30.3	1308	1288
CINMS_B	34.3	-120	0	580	580
W1M3A	43.8	9.1	0	36	1200
glider WHOI	42.4	-70.6	4.21	0-99	73

C. Numerical experiment 1 : Applying untrained models on new data

A first numerical experiment aimed to assess the generalization capability of untrained models currently available in the scientific literature linking the $U10$ wind speed (m/s) to the noise level ($dB\ re\ 1\ \mu Pa^2/Hz$) (Tab III). Untrained models refer to analytical models with parameters from literature as detailed in eqs. (1) to (4). Four models were selected, one is logarithmic (Hildebrand [20]), one is a polynomial cubic model (Nystuen [17]), one is a two-regime linear-quadratic model (Pensieri [18]) and the last one is a composite quadratic-log model (Vagle [8]). The models are shown in Figure 1. Note that the different pre-processing methods to obtain noise level values are not considered here.

[1] https://www.cetus.ucsd.edu/dclde/datasetDocumentation.html
[2] http://www.w1m3a.cnr.it
[3] https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=95
[4] https://www.ndbc.noaa.gov/
[5] https://github.com/Project-OSmOSE/OSEkit

TABLE III: Training metadata for each literature model. Distance corresponds to the distance between the hydrophone and wind speed ground truth. Minimum and maximum wind speed indicates the validity range of each model. Duration refers to the total time span between the start of the first recording and the end of the last recording that contribute to a single input of the model, regardless of any gaps with no recording. $\mathcal{M}_{Hildebrand2021}$ is depth dependant.

Study	Pensieri2015	Vagle1990	Nystuen2015	Hildebrand2021
Ground-truth	Anemometer	Anemometer Vector averaging wind recorder	Anemometer	CCMP v2 model
Ground truth type	U10	U10	U10	U10
Distance (km)	0	0	0	<14
Minimum wind (m/s)	2	3.3	2	0.3
Maximum wind (m/s)	25	15	N/A	32.6
Depth (m)	36	150	200 500	N/A
Frequency (Hz)	8000	8000	8000	8000
Frequency bandwidth (Hz)	1000	1	?	100
Duration (s)	4.5	20.125	4.5	3600

$$\mathcal{M}_{\text{Nystuen2015}} : U = 0.0005 \cdot PSD^3 - 0.0310 \cdot PSD^2 + 0.4904 \cdot PSD + 2.0871 \quad (1)$$

$$\mathcal{M}_{\text{Pensieri2015}} : \begin{cases} U = 0.044642 \cdot PSD^2 - 3.2917 \cdot PSD + 63.016, \\ U = 0.1458 \cdot PSD - 3.146, \\ \text{for} \quad PSD < 38 \text{ dB} \end{cases} \quad (2)$$

$$\mathcal{M}_{\text{Vagle1990}} : U = 10^{\left(\frac{-38.70+\sqrt{38.70^2-4\cdot7.38\cdot(PSD-21.69)}}{-7.38\cdot2}\right)} \quad (3)$$

$$\mathcal{M}_{\text{Hildebrand2021}} : U = 10^{\left(\frac{PSD-10\cdot1.3\cdot\log(0.8)-O_{depth}}{20\cdot0.95}\right)} \quad (4)$$

These models, with literature parameters, were directly used to estimate wind speed in our different datasets without any retraining on the benchmark datasets. PSDs above the 99^{th} quantile were removed in a simple flagging process.

The parameters for $\mathcal{M}_{Hildebrand2021}$ were obtained with the reference curves provided in the study's wind noise model[6]. As these parameters are wind and depth dependent for a given frequency, values for Beaufort wind classes B2-B8 were averaged for final parameter selection. For varying depths, only the parameter linked to a noise level offset was found to change significantly, values are reported in Table IV. For $\mathcal{M}_{vagle1990}$, a depth correction also exists within the model [8], however it required sound speed profile data that was not available for every dataset of the corpus. Selected parameters therefore correspond to a model fitted on noise level data at a depth of 1 m.

[6]https://github.com/jahildebrand/WindNoise.git

TABLE IV: Dataset specific depth offsets for the $\mathcal{M}_{hildebrand2021}$ model.

	BOUSSOLE	SOCAL E	CINMS B	W1M3A	glider WHOI
O_{depth}	49.0	45.4	46.6	48.9	48.7

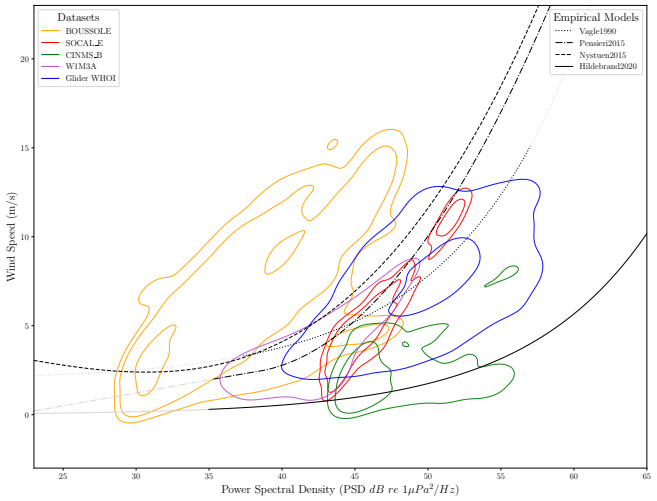


Fig. 1: Density plots of different datasets and curves representing the four models from literature. Black portions represent the wind speeds for which the models were validated.

D. Numerical experiment 2 : PSD shift

In numerical experiment 2, a hypothesis is made that the relationship between a dataset's noise level and wind speed differs solely by a quantifiable PSD shift, due to various factors such as errors in calibration, deployment depth and bathymetry. Associating different offsets to the PSDs is comparable to a normalization of the datasets, as a translation over the PSD axis ensures that any data matches the PSD range from a different deployment. In this numerical experiment, the models used in numerical experiment 1 are considered, modified only by implementing a variable change linked to

the offset S_{off}, parameter that is trained for every dataset, so that $\mathcal{M}_{\text{model}}(PSD)$ becomes $\mathcal{M}'_{\text{model}}(PSD - S_{off})$. In this numerical experiment, we introduce another literature model based on [19]. This model is a two-regime loglinear model $\mathcal{M}'_{cauchy2018}$. Note that the models from [19] and the composite model from [8] were not trained on absolute PSD values. While $\mathcal{M}_{Vagle1990}$ was trained on data with a similar PSD range to that used in this study and could therefore be included in numerical experiment 1, the model defined by Cauchy was not suitable for direct application to our datasets due to differences in its parameterization (third octave levels centered at 3 kHz) and is therefore introduced only in this numerical experiment.

$$\mathcal{M}'_{\text{Cauchy2018}} : \begin{cases} U = \dfrac{1}{0.4 \cdot 10^4} \cdot \left(10^{\frac{PSD - S_{off}}{20}} + 0.2 \cdot 10^4\right) \\ U = \dfrac{1}{1.6 \cdot 10^4} \cdot \left(10^{\frac{PSD - S_{off}}{20}} + 12.5 \cdot 10^4\right) \\ \qquad \text{for} \quad U > 10\ m/s \end{cases} \tag{5}$$

Furthermore, in numerical experiment 1, the noise level value at which the regime change occurs for $\mathcal{M}_{Pensieri2015}$ was based on the given PSD value of 38 $dB\ re\ 1\mu Pa^2/Hz$, as the models are applied under the assumption that no prior information about the ground truth is known. Starting from numerical experiment 2, the regime shifts for the two regime models are based on a wind speed ground truth threshold. Therefore, $\mathcal{M}'_{Pensieri2015}$ is defined as follows :

$$\mathcal{M}'_{\text{Pensieri2015}} : \begin{cases} U = 0.044642 \cdot PSD^2 - 3.2917 \cdot PSD \\ \qquad\qquad + 63.016, \\ U = 0.1458 \cdot PSD - 3.146, \\ \qquad\qquad \text{for} \quad U < 5.5\ m/s \end{cases} \tag{6}$$

E. Numerical experiment 3 : Training models on new data

A third numerical experiment aims to search for optimal performance of the different model types on each benchmark dataset. Each model from numerical experiment 2 is modified so that all parameters become trainable, as follows:

$$\mathcal{M}_{\text{cubic}} : U = a \cdot PSD^3 + b \cdot PSD^2 + c \cdot PSD + d \tag{7}$$

$$\mathcal{M}_{\text{two regime linear-quadratic}} : \begin{cases} U = a_1 \cdot PSD^2 \\ \qquad + b_1 \cdot PSD + c_1, \\ U = a_2 \cdot PSD + b_2 \\ \quad \text{for} \quad U < 5.5\ m/s \end{cases} \tag{8}$$

$$\mathcal{M}_{\text{composite}} : U = 10^{\left(\frac{-b + \sqrt{b^2 + 4 \cdot c \cdot (PSD - a)}}{c \cdot 2}\right)} \tag{9}$$

$$\mathcal{M}_{\text{logarithmic}} : U = 10^{\left(\frac{PSD + 10 \cdot a \cdot \log(8000)}{20 \cdot c}\right)} \tag{10}$$

$$\mathcal{M}_{\text{two regime loglinear}} : \begin{cases} U = a_1 \cdot \left(10^{\frac{PSD}{20}} - b_1\right) \\ U = a_2 \cdot \left(10^{\frac{PSD}{20}} - b_2\right) \\ \qquad \text{for} \quad U > 10\ m/s \end{cases} \tag{11}$$

Model parameters (*a*, *b*, *c*) are freely fitted on the data during the training phase of the two protocols described in Section II-B.

III. Results

A. Numerical experiment 1

Fig 2 demonstrates that applying a untrained model on a new dataset can lead to significant errors. These results were already conveyed by Fig 1 where it is visible that untrained models from literature are not suited to every dataset of the corpus. The MAE varies between 1.12 m/s and 7.58 m/s.

B. Numerical experiment 2

When an trainable offset is added to the models, in numerical experiment 2, the error decreases significantly, with models having errors below 4.2 m/s (Fig 2). Compared to numerical experiment 1, the models show similar performance when evaluated on a specific dataset, with the standard deviation of model performance within each dataset decreasing from 2.06 m/s to 0.93 m/s. However, model performance varies across different datasets, with errors falling below 1 m/s for SOCAL_E, while MAE values exceed 3.5 m/s for Glider WHOI.

TABLE V: MAE score averaged during numerical experiment 3 for both evaluation protocols. Balanced accuracy scores (BAS) for Beaufort classification task.

Metric	S-KCV MAE	T-KCV MAE	S-KCV BAS	T-KCV BAS
Composite	1.52	1.68	0.27	0.25
Cubic	1.50	1.77	0.27	0.19
Logarithmic	1.53	1.70	0.26	0.25
Two regime linear quadratic	1.50	1.68	0.27	0.24
Two regime loglinear	1.34	1.41	0.32	0.29
Mean	**1.48**	**1.65**	**0.28**	**0.24**

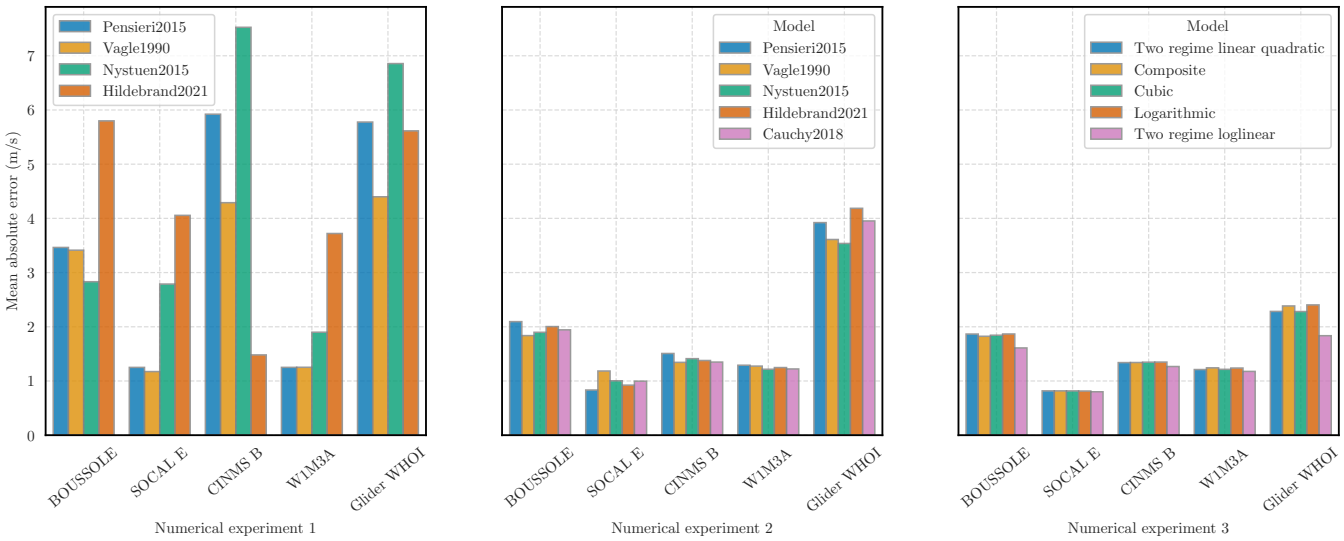
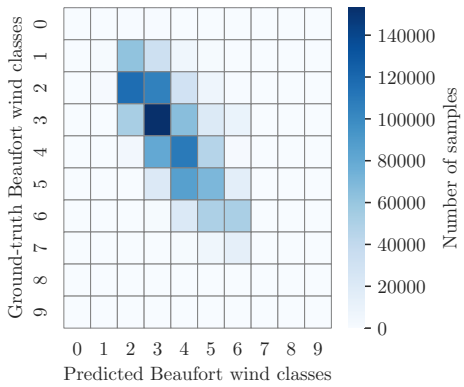
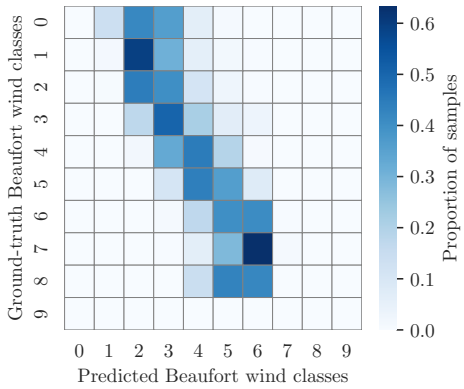


Fig. 2: Performance results for the three numerical experiments using the S-KCV protocol. Results show the MAE for considered untrained, untrained with PSD shift and trained models on each dataset.

C. Numerical experiment 3



(a) Confusion matrix



(b) Normalized confusion matrix

Fig. 3: Confusion matrix (number of samples correctly classified) and normalized confusion matrix (percentage of samples correctly classified or BAS) for the S-KCV protocol in numerical experiment 3 across all datasets.

The third numerical experiment shows that the low complexity models chosen from the literature can be trained to estimate surface wind speed on every dataset and with very similar scores (Fig 2). The number of parameters and nature of the analytical model does not seem to improve accuracy, and predictions are limited by the datasets themselves. The final mean absolute error for all datasets was 1.48 m/s for the S-KCV protocol and 1.65 m/s for the T-KCV protocol (Tab V). Polynomial models performed marginally better than logarithmic ones once trained. The S-KCV numerical experiment outperforms the T-KCV numerical experiment by an average error of 0.17 m/s.

The classification metric in Fig 3 highlights that a vast majority of PSDs are in the correct Beaufort class, with only a few samples being wrong by two classes or more. However, when normalized, results show that B0 and B1 (or wind speeds up to 3.4 m/s) as well as classes upwards from B6 (10.8 m/s and above) are more incorrectly classified.

IV. Discussion

A. Generalization of untrained literature models

Results from the first numerical experiment show that untrained models from literature cannot be generalized to every dataset from the corpus. Different factors will impact the local relationships between PSDs and wind speed. A first factor, intrinsic to the recording setup, is hydrophone calibration. Calibration of acoustic data is indeed challenging and often imperfect [24]. Hypotheses during the pre-processing of audio data such as assuming a linear frequency response, that the hydrophone is perfectly omnidirectional or that the provided laboratory calibration can be generalized to field conditions lead to important noise level shifts. Other important factors of influence relate to the processing pipelines used in the literature, differing in the methods used to compute spectral

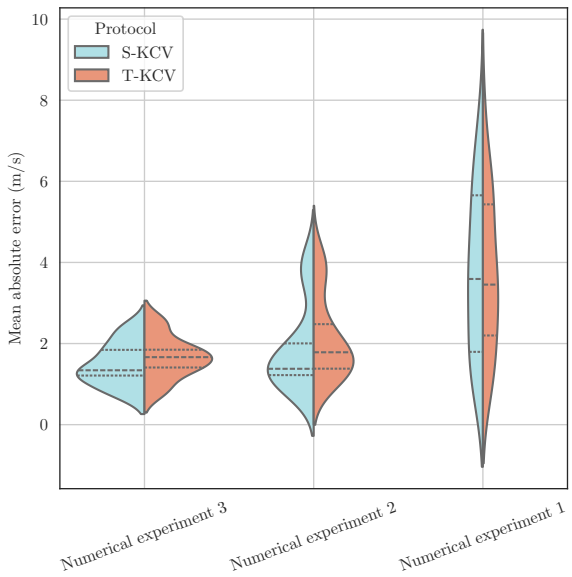


Fig. 4: MAE scores for the two protocols (S-KCV and T-KCV) averaged across all datasets and models for the three numerical experiments.

variables (e.g., third-octave-levels, parameters of PSDs like frequency bandwidth) and the duration of the audio recordings considered. Other factors are related to the local oceanographic and acoustic environments, including bathymetry, biological life, shipping lanes and sound speed profile, that will influence ambient noise in the different sites [23]. This ocean floor noise, or the ambient noise for a given deployment with no wind generated sound or transient sources (ships/marine life/rain etc.), needs to be assessed to obtain accurate wind speed estimations. For instance, while CINMS_B and SOCAL_E are close to each other, the first is at the entry of a channel while the latter is in a more open environment with greater bathymetry, which could explain the difference in average PSD from Fig 1. This prevents current models from being blindly embedded in underwater deployments.

B. Main sources of error

Results show that an important source of error comes from a difference in the noise level range. This shift in PSD between different deployments translates itself through the $1.85\ m/s$ MAE decrease between numerical experiment 1 and numerical experiment 2 (Fig. 4). This represents 50% of the total MAE for untrained models. This error is caused by two main factors : *i)* approximations in the calibration of PSDs already mentioned in IV-A, and *ii)* a PSD offset caused by the instrument's depth. Based on [8], PSDs can be corrected to match surface noise level, depending on the local sound velocity profile. [19] reported an underestimation of $1.3\ dB$ for measurements at a depth of 1000 m. This reinforces the idea that, assuming perfect calibration, a site-specific threshold noise level needs to be known and considered in order to create a universal analytical model [25].

C. Trained model performance

Overall, every model described in the literature can be trained on a new dataset. The discrepency between numerical experiment 2 and numerical experiment 3 of $0.37\ m/s$ (Fig. 4) can mostly be attributed to changes in the relationship between noise level and surface wind speed or changes in frequency response of the different instruments. It accounts for about 10% of the total error. The difference between the T-KCV and S-KCV training can be explained by changing conditions, and varying wind regimes through the timescale and seasonality of individual datasets. For long term deployments, shifts in wind intensity will influence training distribution and create uneven splits. The changing conditions also affect sound propagation through temperature and salinity variations, and the number and diversity of sound sources in the environments such as biological and anthropogenic noise production is also likely to change. Similarly, if the deployment period is too short, temporal splits have higher chances to be uneven and to not contain representative data encompassing high and low wind speeds, thus decreasing the final accuracy.

D. Errors for low and high wind speeds

As illustrated in Fig 3, analytical models are not particularly suitable for low and high wind speeds. For Beaufort scales 0 and 1, the ocean surface is described as calm with some wind-induced ripples, but without foam crests or whitecaps. Under such conditions no wind-generated noise is produced, and provided that no other sound source is present, the PSD corresponds to the floor noise level. Regarding high wind speeds, multiple reasons can explain why the analytical models make poor estimations. First of all, the presence of outliers where the computed PSD is high during low or medium wind speeds is greatly detrimental when fitting the model. In most cases, such occasions - e.g. bioacoustic noise and shipping noise - are more frequent than very high winds and sometimes have similar spectral shapes [18]. This is one main reason why flagging such events is important. Secondly, for high Beaufort wind scales (B6 and B7), although the height of waves increases and the density of foam creases intensifies injecting large number of bubbles into the ocean surface layers, these bubbles absorb sound produced by new breaking waves [25]–[27]. For hydrophones installed below the bubble layer, this absorption causes the sound levels at higher frequencies to be lower than the sound levels acquired at very low wind speed. This phenomenon limits the use of the trained models from numerical experiment 3 for extreme wind events. Such events can however be estimated by training models over different frequencies or if enough data for high wind events are available. While models in this study were trained at $8\ kHz$ which might not be the resonant frequency of the oscillating bubbles for higher winds and is more impacted by bubble attenuation [25], [28], studies have shown that wind speeds during hurricanes can be estimated using frequencies under $100\ Hz$ [29], [30]. Finally, the models were trained on datasets where low wind speeds (B0-B2) and high wind speeds (B6 and above) are scarce, which explains the lower classification

score when classes are normalized. Data augmentation techniques and a weighted loss [19] should be beneficial to the overall model performance. Creating balanced models that can leverage multiple frequencies to predict surface wind speed, or training multiple analytical models over different frequency ranges will better handle low and high wind speed scenarios.

V. Conclusion

Blind operational use of acoustic meteorology models remains unfeasible with low complexity models that only perform well with prior training on data from a similar deployment. The difference in MAE between the untrained models, offset models and trained models suggests that the main source of error comes from calibration errors and the difference in background noise level, and a smaller source of error stems from the variations of the frequency response of ambient noise to surface wind speed. With perfect calibration, spatial and temporal variations of *i)* floor noise and *ii)* wind speed regimes are the main obstacles for the training of a universal model.

Expanding acoustic meteorology models to incorporate diverse observational datasets could enhance their robustness and generalizability. Wind speed estimates from autonomous gliders and BGC-Argo floats provide independent constraints that, when integrated, form a more comprehensive training dataset for machine learning and statistical models. By spanning a wider range of oceanic conditions and acoustic environments, these heterogeneous data sources may reduce overfitting and improve model transferability across regions.

Passive acoustic monitoring also holds potential for refining satellite-based wind speed retrievals, which rely on empirical algorithms that benefit from in-situ validation. Improved acoustic-based wind estimates could strengthen the accuracy of satellite-derived products, bridging gaps in global wind field models. The integration of spaceborne and in-situ acoustic measurements may ultimately enhance weather forecasting and deepen our understanding of air-sea interactions in the climate system.

References

[1] P. Tyack, "Ocean Sound Essential Ocean Variable Implementation Plan". Zenodo, Nov. 08, 2023. doi: 10.5281/zenodo.10067187.

[2] R. Wanninkhof, "Relationship between wind speed and gas exchange over the ocean", Journal of Geophysical Research: Oceans, vol. 97, n. C5, p. 7373-7382, 1992, doi: 10.1029/92JC00188.

[3] D. K. Woolf et al., "Modelling of bubble-mediated gas transfer: Fundamental principles and a laboratory test", Journal of Marine Systems, vol. 66, n. 1, p. 71-91, Jun. 2007, doi: 10.1016/j.jmarsys.2006.02.011.

[4] A. Roobaert, G. G. Laruelle, P. Landschützer, et P. Regnier, "Uncertainty in the global oceanic CO2 uptake induced by wind forcing: quantification and spatial analysis", Biogeosciences, vol. 15, n. 6, p. 1701-1720, Mar. 2018, doi: 10.5194/bg-15-1701-2018.

[5] A. R. Fay et al., "Utilizing the Drake Passage Time-series to understand variability and change in subpolar Southern Ocean pCO2", Biogeosciences, vol. 15, n. 12, p. 3841-3855, Jun. 2018, doi: 10.5194/bg-15-3841-2018.

[6] L. Gregor, J. Shutler, et N. Gruber, "High-Resolution Variability of the Ocean Carbon Sink", Global Biogeochemical Cycles, vol. 38, n. 8, p. e2024GB008127, 2024, doi: 10.1029/2024GB008127.

[7] V. Knudsen O., R. S. Alford, et J. W. Emling, "Underwater Ambient Noise", Journal of Marine Research, vol. 7, p. 410-429, 1948.

[8] S. Vagle, W. G. Large, et D. M. Farmer, "An Evaluation of the WOTAN Technique of Inferring Oceanic Winds from Underwater Ambient Sound", Journal of Atmospheric and Oceanic Technology, vol. 7, n. 4, p. 576-595, Aug. 1990, doi: 10.1175/1520-0426(1990)007¡0576:AEOTWT¿2.0.CO;2.

[9] T. E. Heindsmann, R. H. Smith, et A. D. Arneson, "Effect of Rain upon Underwater Noise Levels", The Journal of the Acoustical Society of America, vol. 27, n. 2, p. 378-379, Mar. 1955, doi: 10.1121/1.1907897.

[10] F. Laville, G. D. Abbott, et M. J. Miller, "Underwater sound generation by rainfall", The Journal of the Acoustical Society of America, vol. 89, n. 2, p. 715-721, Feb. 1991, doi: 10.1121/1.1894631.

[11] B. B. Ma, J. A. Nystuen, et R.-C. Lien, "Prediction of underwater sound levels from rain and wind", The Journal of the Acoustical Society of America, vol. 117, n. 6, p. 3555-3565, Jun. 2005, doi: 10.1121/1.1910283.

[12] J. Yang, S. Riser, J. Nystuen, W. Asher, et A. Jessup, "Regional Rainfall Measurements Using the Passive Aquatic Listener During the SPURS Field Campaign", oceanog, vol. 28, n. 1, p. 124-133, Mar. 2015, doi: 10.5670/oceanog.2015.10.

[13] T. Alsarayreh et L. Zedel, "Quantifying Snowfall Rates using Underwater Sound", Atmosphere-Ocean, vol. 49, n. 2, p. 61-66, Jun. 2011, doi: 10.1080/07055900.2011.573463.

[14] G. M. Wenz, "Acoustic Ambient Noise in the Ocean: Spectra and Sources", The Journal of the Acoustical Society of America, vol. 34, n. 12, p. 1936-1956, Dec. 1962, doi: 10.1121/1.1909155.

[15] D. M. Farmer et D. D. Lemon, "The Influence of Bubbles on Ambient Noise in the Ocean at High Wind Speeds", J. Phys. Oceanogr., vol. 14, n. 11, p. 1762-1778, Nov. 1984, doi: 10.1175/1520-0485(1984)014¡1762:TIOBOA¿2.0.CO;2.

[16] H. C. Pumphrey et J. E. Ffowcs Williams, "Bubbles as sources of ambient noise", IEEE Journal of Oceanic Engineering, vol. 15, n. 4, p. 268-274, Oct. 1990, doi: 10.1109/48.103520.

[17] J. A. Nystuen, M. N. Anagnostou, E. N. Anagnostou, et A. Papadopoulos, "Monitoring Greek Seas Using Passive Underwater Acoustics", Journal of Atmospheric and Oceanic Technology, vol. 32, n. 2, p. 334-349, Feb. 2015, doi: 10.1175/JTECH-D-13-00264.1.

[18] S. Pensieri, R. Bozzano, J. A. Nystuen, E. N. Anagnostou, M. N. Anagnostou, et R. Bechini, "Underwater Acoustic Measurements to Estimate Wind and Rainfall in the Mediterranean Sea", Advances in Meteorology, vol. 2015, p. e612512, Apr. 2015, doi: 10.1155/2015/612512.

[19] P. Cauchy, K. J. Heywood, N. D. Merchant, B. Y. Queste, et P. Testor, "Wind Speed Measured from Underwater Gliders Using Passive Acoustics", Journal of Atmospheric and Oceanic Technology, vol. 35, n. 12, p. 2305-2321, Dec. 2018, doi: 10.1175/JTECH-D-17-0209.1.

[20] J. A. Hildebrand, K. E. Frasier, S. Baumann-Pickering, et S. M. Wiggins, "An empirical model for wind-generated ocean noise", The Journal of the Acoustical Society of America, vol. 149, n. 6, p. 4516-4533, Jun. 2021, doi: 10.1121/10.0005430.

[21] D. Cazau, J. Bonnel, J. Jouma'a, Y. le Bras, et C. Guinet, "Measuring the Marine Soundscape of the Indian Ocean with Southern Elephant Seals Used as Acoustic Gliders of Opportunity", Journal of Atmospheric and Oceanic Technology, vol. 34, n. 1, p. 207-223, Jan. 2017, doi: 10.1175/JTECH-D-16-0124.1.

[22] F. Ingenito et S. N. Wolf, "Site dependence of wind-dominated ambient noise in shallow water", The Journal of the Acoustical Society of America, vol. 85, n. 1, p. 141-145, Jan. 1989, doi: 10.1121/1.397722.

[23] C. M. Duarte et al., "The soundscape of the Anthropocene ocean", Science, vol. 371, n. 6529, p. eaba4658, Feb. 2021, doi: 10.1126/science.aba4658.

[24] A. M. Hurrell et S. Rajagopal, "The Practicalities of Obtaining and Using Hydrophone Calibration Data to Derive Pressure Waveforms", IEEE Trans Ultrason Ferroelectr Freq Control, vol. 64, n. 1, p. 126-140, Jan. 2017, doi: 10.1109/TUFFC.2016.2594770.

[25] R. Vakkayil, H. C. Graber, et W. G. Large, "Oceanic winds estimated from underwater ambient noise observations in SWADE", in OCEANS 96 MTS/IEEE Conference Proceedings. The Coastal Ocean - Prospects for the 21st Century, Sep. 1996, p. 45-51 vol.1. doi: 10.1109/OCEANS.1996.572450.

[26] Z. Zhao, E. A. D'Asaro, et J. A. Nystuen, "The Sound of Tropical Cyclones", Journal of Physical Oceanography, vol. 44, n. 10, p. 2763-2778, Oct. 2014, doi: 10.1175/JPO-D-14-0040.1.

[27] [1] J. Thomson et al., "Surface Wave Development and Ambient Sound in the Ocean", JGR Oceans, vol. 129, n. 12, p. e2024JC021921, Dec. 2024, doi: 10.1029/2024JC021921.

[28] L. Bjørnø, "Chapter 6 - Ambient Noise", in Applied Underwater Acoustics, T. H. Neighbors et D. Bradley, Éd., Elsevier, 2017, p. 363-401. doi: 10.1016/B978-0-12-811240-3.00006-0.
[29] J. D. Wilson et N. C. Makris, "Ocean acoustic hurricane classification", The Journal of the Acoustical Society of America, vol. 119, n. 1, p. 168-181, Jan. 2006, doi: 10.1121/1.2130961.
[30] J. D. Wilson et N. C. Makris, "Quantifying hurricane destructive power, wind speed, and air-sea material exchange with natural undersea sound", Geophysical Research Letters, vol. 35, n. 10, 2008, doi: 10.1029/2008GL033200.

APPENDIX

TABLE VI: Every result (MAE and BAS scores for both evaluation protocols) for the three numerical experiments. Results are grouped by dataset then by model. Please note that depending on the numerical experiment, model name is different.

		Numerical Experiment 1				Numerical Experiment 2				Numerical Experiment 3			
		S-KCV MAE	T-KCV MAE	S-KCV BAS	T-KCV BAS	S-KCV MAE	T-KCV MAE	S-KCV BAS	T-KCV BAS	S-KCV MAE	T-KCV MAE	S-KCV BAS	T-KCV BAS
BOUSSOLE	Two regime linear-quadratic Pensieri2015	3.46	3.46	0.08	0.08	2.10	2.88	0.32	0.27	1.87	1.98	0.23	0.22
	Composite Vagle1990	3.41	3.20	0.12	0.12	1.84	1.97	0.28	0.29	1.82	1.85	0.24	0.30
	Cubic Nystuen2015	2.83	2.72	0.12	0.12	1.90	2.21	0.28	0.27	1.84	1.80	0.23	0.21
	Logarithmic Hildebrand2021	5.80	5.36	0.12	0.12	2.00	2.45	0.29	0.25	1.87	1.92	0.23	0.28
	Two regime loglinear Cauchy2018					1.94	2.48	0.29	0.26	1.61	1.56	0.25	0.28
SOCAL E	Two regime linear-quadratic Pensieri2015	1.25	1.25	0.44	0.44	0.83	1.06	0.44	0.41	0.82	0.91	0.52	0.51
	Composite Vagle1990	1.17	1.12	0.39	0.39	1.18	1.38	0.39	0.18	0.82	0.86	0.52	0.50
	Cubic Nystuen2015	2.79	2.95	0.15	0.15	1.00	1.18	0.40	0.23	0.82	1.42	0.52	0.30
	Logarithmic Hildebrand2021	4.06	3.68	0.17	0.17	0.92	1.10	0.40	0.32	0.81	0.83	0.51	0.53
	Two regime loglinear Cauchy2018					1.00	1.28	0.40	0.23	0.80	0.85	0.47	0.38
CINMS B	Two regime linear-quadratic Pensieri2015	5.92	5.92	0.03	0.03	1.51	1.62	0.15	0.14	1.34	1.38	0.15	0.14
	Composite Vagle1990	4.29	4.31	0.20	0.20	1.34	1.34	0.14	0.14	1.34	1.41	0.14	0.14
	Cubic Nystuen2015	7.53	7.58	0.01	0.01	1.41	1.43	0.14	0.14	1.35	1.42	0.15	0.15
	Logarithmic Hildebrand2021	1.48	1.54	0.15	0.15	1.38	1.43	0.16	0.16	1.35	1.41	0.14	0.14
	Two regime loglinear Cauchy2018					1.35	1.33	0.14	0.15	1.27	1.34	0.28	0.28
W1M3A	Two regime linear-quadratic Pensieri2015	1.25	1.25	0.45	0.45	1.29	1.79	0.35	0.16	1.21	1.66	0.30	0.17
	Composite Vagle1990	1.25	1.42	0.26	0.26	1.27	1.78	0.28	0.18	1.24	1.77	0.31	0.17
	Cubic Nystuen2015	1.90	2.42	0.38	0.38	1.22	1.74	0.30	0.17	1.21	1.74	0.30	0.17
	Logarithmic Hildebrand2021	3.72	3.44	0.09	0.09	1.25	1.80	0.32	0.17	1.24	1.79	0.31	0.17
	Two regime loglinear Cauchy2018					1.22	1.67	0.30	0.17	1.18	1.48	0.36	0.28
Glider WHOI	Two regime linear-quadratic Pensieri2015	5.78	5.78	0.14	0.14	3.92	3.94	0.13	0.20	2.28	2.48	0.15	0.15
	Composite Vagle1990	4.40	4.01	0.19	0.19	3.61	3.76	0.13	0.15	2.38	2.52	0.13	0.14
	Cubic Nystuen2015	6.86	5.66	0.11	0.11	3.54	3.80	0.15	0.21	2.28	2.48	0.15	0.15
	Logarithmic Hildebrand2021	5.61	5.86	0.12	0.12	4.19	4.34	0.14	0.16	2.40	2.53	0.12	0.14
	Two regime loglinear Cauchy2018					3.95	3.72	0.14	0.16	1.84	1.84	0.22	0.22