Learning-Based Temporal Estimation of In-Situ Wind Speed From Underwater Passive Acoustics

Matteo Zambra, Dorian Cazau, Nicolas Farrugia, Alexandre Gensse, Sara Pensieri, Roberto Bozzano, and Ronan Fablet, Senior Member, IEEE

Abstract—Wind speed retrieval at the sea surface is of primary importance for scientific and operational applications. Besides weather models, in-situ measurements and remote sensing technologies, especially satellite sensors, provide complementary means to monitor wind speed. As sea-surface winds produce sounds that propagate underwater, underwater acoustics recordings can also deliver fine-grained wind-related information. Whereas model-driven schemes, especially data assimilation approaches, are the state-of-the-art schemes to address inverse problems in geoscience, machine learning techniques have become more and more appealing to fully exploit the potential of observation data sets. Here, we introduce a deep learning approach for the retrieval of wind speed time series from underwater acoustics possibly complemented by other data sources such as weather model reanalyses. Our approach bridges data assimilation and learning-based frameworks that benefit from both prior physical knowledge and computational efficiency. Numerical experiments on real data demonstrate that we outperform the state-of-the-art data-driven methods with a relative gain of up to 16% in terms of root-mean-squared error (RMSE). Interestingly, these results support the relevance of the time dynamics of underwater acoustic data to better inform the time evolution of wind speed. They also show that multimodal data, here underwater acoustics data combined with European Center of Medium-Range Weather Forecast (ECMWF) reanalysis data, may further improve the reconstruction performance, including the robustness with respect to missing underwater acoustics data.

Index Terms—Data assimilation, deep learning, geophysical signal processing, underwater acoustics.

I. INTRODUCTION

Wind speed monitoring is of key interest for a wide range of applications and domains including climate and atmosphere science, meteorological modeling, and routing applications. Besides in-situ measurements, remote sensing technologies have long been explored for the estimation of wind speed [1], [2], [3], [4], in particular through satellite sensors. This class of techniques allows for a high-resolution and weather-independent sensing of the sea surface (like in the case of synthetic aperture radar (SAR) imagery [5]). However, the revisit time of satellite SAR sensors may be as large as 10 days, which cannot inform rapidly-evolving wind speed patterns from the hourly scale to the daily one. This has motivated the exploration of other remote sensing technologies such as underwater acoustic sensors [6], [7] to deliver complementary low-cost and nonintrusive indirect measurements of the sea-surface wind speed. Such sensors can be deployed for a relatively long period of time, typically from weeks to years and can sample underwater noise continuously with a high time resolution.

The seminal work of Nystuen [8] prompted the development of acoustic meteorology, a discipline that aims to reconstruct above-surface meteorological phenomena (such as wind speed and rainfall) given ocean ambient noise. The instruments typically used to capture underwater ambient noise are fixed hydrophones [9], [10], mobile acoustic platforms as ARGO profilers [11], [12], and finally free-ranging bio-logged marine mammals [13]. The developments of acoustic meteorology made it possible to move toward an operationalization of passive acoustic measurements for geophysical phenomena, for example, the estimation of wind speed starting from underwater noise measurement [7], [12], [14]. Recently, machine learning techniques led to the further improvement in the prediction performance [15]. The appeal of machine learning approaches stems from their computational efficiency, the availability of open-source libraries, and the ability to skip handcrafted feature engineering steps.

From a methodological point of view, wind speed retrieval from acoustic data can be regarded as an inverse problem [16], which classically relies on state-space formulation when considering time-evolving processes. However, the complexity of the physical laws regulating the wind speed dynamics and the variability of the water column conditions that affect the underwater sound propagation prevents the design of a classic model-driven inversion scheme. However, it also motivates the exploration of purely data-driven schemes in previous works [15], [17], [18], which do not explicitly account for time-related patterns. Interestingly, recent studies have introduced physics-informed end-to-end learning schemes for inverse problems with time

Manuscript received 19 August 2022; revised 9 May 2023; accepted 13 June 2023. Date of publication 24 August 2023; date of current version 13 October 2023. This work was supported in part by the AI Chair Oceania under ANR Grant ANR-19-CHIA-0016 and in part by the industrial partnership with Naval Group. (Corresponding author: Matteo Zambra.)

Associate Editor: T. Ross.

Matteo Zambra, Nicolas Farrugia, and Ronan Fablet are with the IMT Atlantique Bretagne-Pays de la Loire and Lab-STICC CNRS URM 6285, Brest, France (e-mail: matteo.zambra@imt-atlantique.fr; nicolas.farrugia@imt-atlantique.fr; ronan.fablet@imt-atlantique.fr).
Dorian Cazau is with the ENSTA Bretagne and Lab-STICC CNRS URM 6285, Brest, France (e-mail: dorian.cazau@ensta-bretagne.fr).
Alexandre Gensse is with the Naval Group, 83190 Ollioules, France (e-mail: alexandre.gensse@naval-group.com).
Sara Pensieri and Roberto Bozzano are with the National Research Council of Italy, Institute for the Study of the Anthropic Impacts and Sustainability in the Marine Environment (CNR-IAS), 16149 Genova, Italy (e-mail: sara.pensieri@cnr.it; roberto.bozzano@cnr.it).

Digital Object Identifier 10.1109/JOE.2023.3288970

1558-1691 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
processes [19]. Here, we extend this line of work to the prediction of sea-surface wind speed from underwater acoustics measurements. We exploit the 4DVarNet end-to-end architecture introduced in [20], which relies on a variational data assimilation formulation. More specifically, it represents the underlying physics through trainable neural network modules and addresses the associated optimization procedure through a trainable solver. While making explicit the modeling and inversion of time processes, we can learn all trainable components from data so that the underlying physical representation is optimized for the considered case study. We emphasize the importance of the models’ capability to effectively represent temporal dynamics in geophysical problems. A primary aspect in Earth sciences is dynamics, and its effective modelization is as valuable as the accurate cause-to-effect mapping. Otherwise phrased, understanding and modeling how in-situ wind speed evolves is as important as accurately estimating the value of in-situ wind speed given the underwater passive acoustic (UPA) datum at a given time. The 4DVarNet is particularly suited for this application, given the time-dependent state-space statement of the underlying physical problem. We also explore a multimodal machine learning approach by incorporating different data sources, namely underwater acoustics and reanalyzed wind speed values available in the ERA-interim database [21]. Multimodality allows us to incorporate heterogeneous information from different data sources [22].

The rest of the article is organized as follows. Section II introduces the relevance of the time dependence of wind speed and provides a suitable inverse problem formulation thanks to the state-space representation [23]. Section III presents our dataset and the multimodal approach for this case study. Section IV details the 4DVarNet end-to-end architecture and how it is applied to the present case study. Section V presents the results obtained and Section VI critically discusses these results.

II. PROBLEM STATEMENT

State-of-the-art methods for the retrieval of sea-surface wind speed from underwater acoustics data rely on machine learning schemes [15]. They state the reconstruction problem as a regression problem, formally

\[ y_t \mapsto u_t \]  

where \( y_t \) refers to the acoustic spectrum at time \( t \) and \( u_t \) is the associated in-situ wind speed.

By contrast, the resolution of inversion problems in geoscience for time-related processes generally relies on a data assimilation formulation [24], [25], [26], [27]. It explicitly accounts for the underlying temporal dynamics of the process of interest through a state-space model

\[ \begin{aligned} \dot{x}(t) &= \mathcal{M}(t, x(t)) \\ y(t) &= \mathcal{H}(t, x(t)) \end{aligned} \]  

where \( x \in X \) represents the state variable defined on the support \([0, T]\) and \( \mathcal{M} \) is a given mathematical model of the physical dynamics. The variable \( y \in Y \) represents a measurement of \( x \) through observation operator \( \mathcal{H} \). When dealing with irregularly-sampled observation data, the observation operator \( \mathcal{H} \) involves a masking operator. Besides the dynamical model \( \mathcal{M} \), we can define the associated flow operator \( \Phi \) as the one-step-ahead predictor of the state \( x(t + \Delta t) \) based on the integration of the model \( \mathcal{M} \) from time \( t \)

\[ \Phi(x, t + \Delta t) = x(t) + \int_t^{t+\Delta t} \mathcal{M}(x, t) \, dt. \]  

The numerical implementation of the flow operator \( \Phi \) involves a numerical integration scheme, such as the Euler and Runge–Kutta explicit integration schemes [28], [29] using time discretization. Given the above state-space formulation, we can formulate a data assimilation problem to reconstruct the state sequence \( x \) from the observation data \( y \) as the minimization of the following variational criterion:

\[ U_{\Phi}(x, y, \Omega) = \lambda_1 \sum_{t=0}^{T} \| x(t) - y(t) \|_{\Omega}^2 + \lambda_2 \sum_{t=0}^{T} \| x(t) - \Phi(x; t) \|_2^2 \]  

where \( \Omega \) is the spatiotemporal domain in which the observations are defined, and \( \lambda_1 \) and \( \lambda_2 \) are tunable weights. To state the previous equation more compactly, we may use the vector notation

\[ U_{\Phi}(x, y, \Omega) = \lambda_1 \| x - y \|_{\Omega}^2 + \lambda_2 \| x - \Phi(x) \|_2^2. \]  

Then, the optimization problems read as follows:

\[ \hat{x} = \arg \min_{x \in X} U_{\Phi}(x, y, \Omega). \]  

We may emphasize that this variational cost comprises two terms: 1) a data fidelity term to assess the agreement between the reconstructed states and the observation data and 2) a prior term to constrain the reconstructed states to match the dynamics associated with model \( \mathcal{M} \). While a broad literature describes data assimilation algorithms to solve for the above minimization problem, a key issue for an application to in-situ wind speed retrieval is the ability to define and parameterize the underlying state-space formulation. Here, we may emphasize that physical models for the underwater acoustics dynamics as well as the relationship between underwater acoustics data and wind speed [14], [30] appear too complex to explore a fully model-driven formulation.

Interestingly, recent studies have explored how to bridge deep learning schemes and data assimilation methods [31], [32], [33], [34], [35], [36] as a way to combine physically-sound formulations with the computational efficiency and the versatility of deep learning frameworks. As reported in [20], we have a generic end-to-end deep learning scheme for time-related inverse problems, which explicitly relies on a variational data assimilation formulation.

III. DATA

Data used in this case study are underwater ambient noise spectra, synthetic reanalyses of wind speed provided by the European Center of Medium-Range Weather Forecast (ECMWF) and in-situ measurements of the wind speed at 10 m above the sea surface. ECMWF reanalyses are publicly available in the ERA-interim database [37] and, UPA data and in-situ wind speed are sampled on the W1M3A marine observatory. A brief
The atmospheric model is coupled to an ocean-wave model with assimilation of a large body of different in situ and satellite data. The wind speed reanalyses come from the ERA-Interim data set, which is a reanalysis of the physical evolution of the meteorological variables involved. The wind speed reanalyses are obtained by combining the available observations and the prior knowledge of the physical model [21]. Since ECMWF wind is an estimation obtained with a numerical model, it is smoothed but implicitly carries information about the physical evolution of the meteorological variables involved. The wind speed reanalyses come from the ERA-Interim data set, based on the global atmosphere model reanalysis developed at the ECMWF. All these global reanalyses are obtained with the assimilation of a large body of different in situ and satellite data. The atmospheric model is coupled to an ocean-wave model with a $1.0^\circ \times 1.0^\circ$ latitude and longitude grid. A detailed description of the ERA-Interim product archive can be found in [21] and [37].

B. W1M3A Observation System

The observation site is the Western 1-Mediterranean Moored Multisensor Array (W1M3A), which is part of EMSO and ICOS networks of European research infrastructures. It is located in the Gulf of Genoa (Italy) at a distance of 80 km from the coast. The W1M3A system is composed of the ODAS Italia 1 spar buoy and a subsurface moored component. Pensieri et al. [30] provide a detailed explanation of the W1M3A observation site. Underwater noise data are collected with a UPA Listener (UPAL).² A detailed description of the functioning of UPAL is provided in Zuba et al. [6], Nystuen et al. [7], and Pensieri et al. [30]. For the sake of context, we provide here some insights of UPAL data collection and preprocessing. UPAL is equipped with a hydrophone but, due to hardware constraints (battery life and memory), it is not possible to sample noise continuously during the whole instrument duty cycle. Noise is sampled for 4.5 s at 100 kHz and then processed to obtain a spectral representation of the signal having 64 frequency bands. The data used in this case study are, indeed, time series of such spectra. UPAL electronics is endowed with a recognition algorithm that classifies ambient noise sources. The duty cycle of the instrument is slightly adapted based on such rough classification, and on average, a recording is acquired every 5 min. The interested reader may refer to the work in [7] for a comprehensive explanation of the UPAL functioning. The frequency bands associated with the wind speed signature are those related to 8 and 20 kHz. Nevertheless, we choose to retain all the frequency bands since one advantage of deep learning modeling stems from the minimal handmade feature engineering required on input data.

Fig. 1. Data set quantitative and qualitative characteristics. (a) Different UPA spectra according to the associate in-situ wind speed, in Beaufort scale. (b) Scatter plot between in-situ and ECMWF wind speed. (c) Probability densities of in-situ wind speed. This is to quantify the values involved.

A. ECMWF Wind Speed Values

ECMWF wind speeds come from model reanalyses.¹ These reanalyses are obtained by combining the available observations and the prior knowledge of the physical model [21]. Since ECMWF wind is an estimation obtained with a numerical model, it is smoothed but implicitly carries information about the physical evolution of the meteorological variables involved. The wind speed reanalyses come from the ERA-Interim data set, based on the global atmosphere model reanalysis developed at the ECMWF. All these global reanalyses are obtained with the assimilation of a large body of different in situ and satellite data. The atmospheric model is coupled to an ocean-wave model with a $1.0^\circ \times 1.0^\circ$ latitude and longitude grid. A detailed description of the ERA-Interim product archive can be found in [21] and [37].

C. Temporal Resolutions

The underwater acoustics data are sampled from 2011-06-17 at 00:50 to 2013-09-06 at 18:50 almost continuously with an hourly resolution, except a period of time between 2013-04-26 and 2013-06-06 in which no observations are available. In-situ wind speed data are available from 2011-06-17 at 00:50 to 2013-09-06 at 18:50, except for a time window from 2012-11-06 to 2013-06-06 where no data are available. In-situ wind speed data are provided as the hourly average of the monitored wind speed values. In an open sea context, such as the W1M3A system, there is almost always nonnegligible wind speed, except for some short periods of time [30]. The hourly resolution is motivated by the fact that UPA and in-situ wind speed values are available as hourly averages. In the previous section, we mentioned the functioning principles of the UPA measurement instrument. Recall that the instrument records underwater noise

¹In the data assimilation literature, a “reanalysis” is the reconstruction of a given process over a given time horizon from a series of observations over the same period of time based on the assimilation of these observation data sources in a dynamical model.

²Not to be confused with UPA. The acronym UPAL refers to the instrument. UPA refers to the data that this instrument provides.
for a time window of 4.5 s, as long as no acoustic sources other than wind speed are present. This time period should not be confused with the temporal resolution of our data set. For a given hour, the noise samples recorded are averaged and this defines the temporal resolution. Time averaging does imply a smoothing of the information contained in the raw data. However, we assume that such a temporal resolution does suffice to retain the dynamical information of wind speed evolution. ECMWF wind speed values are available from 2011-06-01 00:00:00 to 2019-06-30 23:00. Note that this latter data modality does not suffer from the presence of missing data, unlike UPAL and in-situ wind since it is the output of an operational model. UPAL and in-situ wind speed time series, on the other hand, may have one or more missing time steps of observation, due to many reasons, including among others instrument failure or maintenance operations.

### D. Preprocessing Scheme

Our first preprocessing step consists in colocating UPAL, ECMWF, and in-situ wind time series, according to their respective time resolutions. If one time step has only UPAL and/or ECMWF but has not an in-situ wind speed value associated, this time step is removed from the overall data set. If one time step has no UPAL and/or ECMWF but has the in-situ value, it is kept since it may be a proof of robustness in case of missing data in our time series. In other words, in the 100% of the time steps of the data set temporal extent, UPAL and ECMWF data are validated by their respective in-situ wind speed value, but the opposite is not true. To complete the data clean-up, we remove all the observation days that do not have full time series of 24 in-situ wind speed values, because we want the data set to be divisible in time series of length 24 (one day from 00:00 to 23:00). This results in keeping about 98% of the original data set. Thus, by doing, we obtain a collection in the temporal order of 14088 triples constituted by a 64-dimensional UPAL vector (each UPA spectrum has 64 frequency bands), a 1-D ECMWF wind speed value, and a 1-D in-situ wind speed value. For our analyses, we need two versions of the same data set, one time-independent version, and one time-dependent version. The time-independent version is simply the collection of 14088 UPAL, ECMWF, and in-situ wind triples. The time-dependent version of the data set, two choices are possible. The first would group the time-independent version in a series of 24 h according to the hours. Each series would start at 00:00 and end at 23:00. Thus, by doing, the overall number of time series available would be 14008/24 = 587. A second choice could be to randomly extract 2000 time series of length 24 from the time-independent version. In this case, the start/end hour of the series is irrelevant. We choose the second alternative since 14088 samples are sufficient to fit, validate, and test a regression model, but 587 samples (considered to be time series) may not suffice to fit, validate, and test the 4DVarNet model since it comprehends two deep network parameterized models (the dynamical prior and the gradient solver). Note that this version of the data set comprehends 2000 × 24 single UPAL, ECMWF, and in-situ wind triples, unlike the 14088 instances of the time-independent version. But in the time-dependent case, the elemental unit of the data set is the 24-h time series and not the single hour. In other words, a time-independent model takes one data triple, referred to one given hour, and returns the in-situ wind for that hour. The time-dependent model takes as input the time series obtained as described and returns a time series of 24 in-situ wind speed values.

### IV. PROPOSED METHOD

This section details the proposed end-to-end deep learning scheme based on a variational data assimilation formulation for the retrieval of in-situ wind speed from multisource data, namely underwater acoustics and ECMWF data.

#### A. Proposed Variational Data Assimilation Model

In Section II, we introduced the formal statement of the problem. Recall that in a variational data assimilation problem, one has access to some observations $y$ and wants to reconstruct their state variable $x$. Here, the state variable is assumed to be the concatenation of the different available data modalities. We consider in the following, the outputs of single-modal and multimodal versions of the proposed 4DVarNet framework. Assume that $\alpha$ indicates the UPA or UPA with ECMWF modality and $\beta$ the in-situ wind speed modality. In the single-modal version, the state variable is the concatenation of UPAL and in-situ wind speed, $x \in \mathbb{R}^{{N}_\alpha + {N}_\beta}$, and the observable part are represented by UPAL data, $y \in \mathbb{R}^{{N}_\alpha}$ with $N_\alpha = 64$ and $N_\beta = 1$. In the multimodal version, the observable part is composed of the UPAL and ECMWF data, and the state variable is still considered to be the concatenation of UPAL, ECMWF, and in-situ wind speed, i.e., with $N_\alpha = 65$ and $N_\beta = 1$. Since we prepare these data as time series, the observation effectively used in the model will be a batch of vectors of shape $(24, N_\alpha)$. Given each time series of 24 UPAL and ECMWF instances, the model will return a time series of 24 wind speed values.

Given these definitions of the state variable and of the observations, we can consider the following observation operator $\mathcal{H}$:

$$
y = [y_\alpha, 0 \times x_\beta] = \mathcal{H}([x_\alpha, x_\beta]).$$

(7)

The square brackets symbolize a simple concatenation operation. The quantity $x_\alpha$ represents the part of the latent variable hosting the acoustics and ECMWF information and $x_\beta$ represents the part of the state variable hosting in-situ wind information. The observation operator has as an argument the concatenation of the information related to observable quantities (marked with $\alpha$) and in-situ wind speed (marked with $\beta$) and returns the only measurable information $y_\alpha$, i.e., related to UPAL and ECMWF. This observation operator simply states that no direct observation of the in-situ wind speed is available. The dependence between the in-situ wind speed and the observed variables derives from the parameterization of the prior operator $\Phi$. Rather than exploring an explicit ODE-based parameterization as in model-driven data assimilation schemes [25], we rely on neural autoencoder architectures as in [19] and [20]. In such architectures, the inputs and outputs share the same shape. Autoencoder architectures [38] have been widely used for denoising, reconstruction, and simulation tasks [39]. They generally exploit a latent lower-dimensional representation of
the input data [39]. This property seems appealing here to enforce the assumption of some underlying latent space jointly encoding sea-surface wind speed and available observation data.

Here, we chose to use a 1-D convolutional autoencoder architecture (Conv-AE) [40]. This is motivated by the fact that convolutional networks can leverage trainable convolution operators to model translation-invariant features throughout data examples. Our encoder is composed of two 1-D convolutional layers, having input–output shapes of $(N_x, N_y, 128)$ and $(128, 20)$. After the first layer, a leaky rectified linear unit nonlinear activation function is placed, and its negative slope is set to $10^{-1}$. The decoder has the same structure, but in reverse order, and the same nonlinear activation functions are used after the two layers. All convolution layers involve a zero-padding and a kernel dimension of 3. The number of channels, 128 and 20, respectively, was set empirically from cross-validation experiments. The architecture of the autoencoder parameterization of $\Phi$ is shown in Fig. 2(a).

**B. Associated Trainable Solver**

Within the proposed approach, the reconstruction of the state variable relies on the minimization of variational cost, as stated in (6). We solve this optimization problem using another neural network model included in the end-to-end architecture. This trainable gradient solver, referred to as $\Gamma$, exploits a convolutional long-short term memory (LSTM) network [41], the latter being particularly suited for time-series modeling [42], and optimizer learning [43]. Overall, the end-to-end architecture implements a predefined number of iterations of the following iterative rule:

$$
\begin{align*}
\phi^k &= \nabla_x U_{\phi}(x^k, y) \\
\beta^k &= \beta^k + \Delta \beta^k
\end{align*}
$$

where $\Gamma$ is a linear layer to map the latent state $h$ of the LSTM cell to the state space. The first row of (8) represents the computation of the gradient of the variational cost w.r.t. the state variable $x$. This variational cost is evaluated with the version of the state variable as computed in the previous iteration. $x^k$ in the first row is different from $x^{k+1}$ in the second row as the former is the state variable computed at iteration $k$ and the latter is the update of the state variable for the iteration $k + 1$. In our implementation, the LSTM cell dimension was set to 100, again based on numerical experiments. We may emphasize that these latent states of the LSTM cell, namely $h$ and $c$, are referred to with the $k$ superscripts, which denotes the $k$th iteration of the gradient solver. A typical choice for the total number of iterations ranges from 5 to 10. This choice derives both the computational complexity of the underlying automatic differentiation [44] as well as the ability of this gradient-based iterative update to converge with only very few gradient-based steps. The architecture of the gradient solver $\Gamma$ is reported in Fig. 2(b). Fig. 2(c) represents the flow diagram of the 4DVarNet framework.

**C. Learning Scheme**

Overall, the training procedure involves two nested gradient descents: 1) an inner minimization of variational cost (5) w.r.t. state variable $x$ and 2) an outer minimization of the training loss w.r.t. model parameters, especially the parameters of models $\Phi$ and $\Gamma, \Theta$. The objective function for the former has been already discussed. Let $\psi_\phi$ be the 4DVarNet end-to-end system and $\hat{x} = \psi_\phi(y)$ the reconstructed state variable. Then, the reconstruction error can be evaluated as

$$
L(x, y) = \lambda_d \|y_{\alpha} - \hat{x}_{\alpha}\|^2 + \lambda_p \|y_{\beta} - \hat{x}_{\beta}\|^2. 
$$

In this equation, $\Omega_{\alpha}$ and $\Omega_{\beta}$ are domain masks that account for missing data. These masks may be simple binary matrices that represent the data sparsity pattern. Parameter optimization is achieved with two different instances of the Adam algorithm [45]. For both the Conv-AE and the LSTM solver, the learning rate is set to $10^{-3}$ and the weight decay to $10^{-5}$. The weights $\lambda_d$ and $\lambda_p$ are set to 0.5 and 1.5 for the Conv-AE and 4DVarNet experiments, respectively. These weights were tuned empirically. The weight on wind prediction is greater because...

---

**Fig. 2.** (a) and (b): Visualizations of the flow operator $\Phi$ parameterized with a neural autoencoder and the trainable gradient solver $\Gamma$ parameterized with an LSTM network, respectively. (c): Flowchart of the 4DVarNet optimization procedure.
it is the term associated with the task of major interest. The full end-to-end architecture $\Psi_\Theta$ is trained for 200 epochs, with no early stopping criteria. Our learning protocol consists of two consecutive steps. First, we perform a full training procedure with five gradient iterations for the 4DVarNet scheme and save the best model based on validation loss. Note the best model is not necessarily the one at the last epoch of the training procedure. Then, this best model is further trained through another full training procedure, this time using 10 gradient iterations.

V. RESULTS

This section reports the numerical experiments carried out to assess the performance of the proposed approach with respect to state-of-the-art data-driven methods.\(^3\) We first detail our evaluation setting and the benchmarked models. We then report and discuss the reconstruction performance for three case studies for the reconstruction of in-situ wind speed 1) using only underwater acoustics data, 2) using underwater acoustics and ECMWF data, 3) when dealing with random gaps in the underwater acoustics data, again in a multimodal UPA and ECMWF data set case.

A. Evaluation Framework

Our evaluation procedure follows the one reported in [15]. We chose as evaluation metrics the root-mean-squared error (RMSE) between reconstructed and in-situ wind speed values. We evaluate this RMSE on the data period from 2011-06-18 00:00 to 2011-08-07 23:00 (50 days), whereas the data period from 2011-08-08 00:00 to 2013-09-05 23:00 is used as an independent training data set, except for the period between 2011-08-25 00:00 and 2011-10-14 23:00 (50 days), which is used as the validation set. From the 50-day time series with an hourly time resolution, we extract all associated one-day time windows, which result in a data set of $(50-1) \times 24 = 1176$ 24-h samples.

For benchmarking purposes, our numerical experiments involve machine learning schemes proposed in [15] as well as state-of-the-art neural network architectures. The latter has been chosen in accordance with the parametrization of 4DVarNet scheme. Overall, the first category of methods involves machine learning schemes, referred to as time-independent, which predict in-situ wind speed at one time step from the underwater acoustics data spectrum at the same time step, as stated in (1). The multimodal approach is not implemented for this class of models. This category includes the following.

1) **CatBoost** [46]: A gradient-boosting algorithm [47], [48], which can manage effectively categorical features. CatBoost uses the loss function and the root-mean-square error as the loss function.

2) **Random Forest** [49]: An ensemble of decision trees, either trained for classification or regression. Random Forest has been set up with the maximum depth of the tree to 10, the number of features to consider when splitting equals the actual number of features, the minimum number of samples required for leaf nodes to 1, and the minimum number of samples required to split an internal node to 2. The number of estimators is set to 100.

3) **FC-AE**: This fully-connected autoencoder (FC-AE) architecture comprises a fully-connected encoder composed of two linear layers with input and output shapes of $(N_\alpha + N_\beta, 128)$ and $(128, 20)$, respectively. After the first layer, a leaky rectifier linear unit with a negative slope of 0.1 is applied. The decoder has the same architecture but is reversed, with the same nonlinearity applied after all layers. The learning rate is set to $10^{-3}$ and the weight decay is set to $10^{-6}$. The weights of the loss terms are chosen to be 0.5 and 1.5. The intermediate and latent dimensions, learning rate and weight decay, and finally, the loss terms weights were set empirically using cross-validation.

The second category of machine learning methods, referred to as time-dependent methods, predicts a time series of in-situ wind speed from a time series of underwater acoustics data and may benefit from time-related features. This category comprises the following schemes.

1) **Fully-connected autoencoder**: The FC-AE parametrization as described above has been reused in the time-dependent configuration. There are no differences in the architecture as the input size does not change, except for the temporal dimension, which is neglected in the time-independent setting.

2) **Conv-AE-UPA**: This convolutional autoencoder architecture refers to the parameterization of operator $\Phi$ in (5) for the proposed 4DVarNet scheme. Here, we use this architecture to train a direct inversion scheme to map input data $y$ to a reconstructed state $x$. We may point out that this direct inversion scheme can be regarded as a single iteration of a fixed-point iterative solver for the minimization of variational cost (5).

3) **Conv-AE-UPA+ECMWF**: This architecture refers to an extension of the previous one when the input data $y$ includes both UPA and ECMWF wind speeds. Besides, the reconstruction $x$ also comprises UPA, ECMWF, and in-situ wind speed states.

4) **4DV arNet-UPA**: Using the Conv-AE as previously described and the trainable solver as detailed in Section IV-B, the first 4DV arNet configuration accounts for the observations $y$ of UPA data only. The output $x$ is a concatenation of UPA reconstructions and in-situ wind speed predictions.

5) **4DV arNet-UPA+ECMWF**: In this second case, the architecture of the 4DV arNet architecture is the same, except for the observations and the state variable. The input data comprehend both UPA and ECMWF wind speeds.

The loss function used to train the deep models is a simple mean-squared error, formulated in (9). For evaluation on the test set, the RMSE is instead used. Since the task studied is the reconstruction of wind speed time series, we consider performance metrics based on the in-situ wind speed data considered as ground truth. To assess uncertainties and have reliable results, 10 runs are performed on a machine equipped with 3 Nvidia Quadro RTX 8000 units. Each of these units has a TU102 graphical processor operating at a frequency of 1395 MHz and
TABLE I
RMSE METRICS FOR TIME-DEPENDENT DATA SET AND MODELS

| Model          | RMSE | Mean ± std | n-Median | Metrics in m s⁻¹ |
|----------------|------|------------|----------|------------------|
| CatBoost       | 0.95 | –          | –        | 0.95             |
| Random Forest  | 0.97 | –          | –        | 0.97             |
| FC-AE          | –    | 0.98 ± 0.03| 0.95     | –                |

TABLE II
MISSING DATA, 4DVARNET-UPA+ECMWF 10 ITER

| Missing data | Mean ± std | n-Median | η [%] |
|--------------|------------|----------|-------|
| 10           | 0.85 ± 0.02| 0.80     | 15.8  |
| 20           | 0.87 ± 0.02| 0.81     | 14.7  |
| 30           | 0.90 ± 0.02| 0.83     | 12.6  |
| 40           | 0.92 ± 0.02| 0.83     | 12.6  |
| 50           | 0.98 ± 0.02| 0.89     | 6.3   |
| 60           | 1.01 ± 0.02| 0.91     | 4.2   |
| 70           | 1.07 ± 0.03| 0.96     | −1.1  |
| 80           | 1.17 ± 0.03| 1.08     | −13.7 |
| 90           | 1.27 ± 0.03| 1.21     | −27.3 |

TABLE III
RESULTS FOR SINGLE AND MULTIMODAL SETTINGS

| Model                        | Mean ± std | n-Median | η [%] |
|------------------------------|------------|----------|-------|
| FC-AE                        | 0.97 ± 0.04| 0.92     | 3.2   |
| Conv-AE-UPA                  | 0.94 ± 0.04| 0.88     | 7.4   |
| 4DVarNet-UPA 5 iter          | 0.91 ± 0.02| 0.87     | 8.4   |
| 4DVarNet-UPA 10 iter         | 0.89 ± 0.04| 0.84     | 11.6  |
| Conv-AE-UPA+ECMWF            | 0.88 ± 0.02| 0.83     | 12.6  |
| 4DVarNet-UPA+ECMWF 5 iter    | 0.85 ± 0.03| 0.81     | 14.7  |
| 4DVarNet-UPA+ECMWF 10 iter   | 0.84 ± 0.02| 0.80     | 15.8  |

The bold text expresses the best experiment in terms of reconstruction root mean squared error (RMSE) and relative gain.

has a 48-GB memory size. In the following tables, two columns are present. The first column named “Mean ± std” reports the average RMSE and the quartiles over the 10 runs. To compare our framework with ensemble models previously discussed, another evaluation strategy is used. Each of the 10 models trained can produce a reconstruction of the wind speed sequence given test data. Similarly to what is done in bagging [50], we chose to compute the median of the wind speed values reconstruction as aggregated output. We express this aggregated output formally as

\[
\hat{x}_{\text{median}} = \text{Median}(\{ \hat{x}^n_n ; 0 \leq n < N_{\text{runs}} \}).
\]  

(10)

The n-median metric is defined as the RMSE between this quantity and the ground truths. To quantify the improvement of the proposed class of models with respect to the baselines, we may define a relative gain metric. Call \( p_b \) and \( p_i \) the baseline and improved performance metrics, chosen to be the ensemble n-median scores. Then, define the relative gain as

\[
\eta = \left(1 - \frac{p_b}{p_i}\right) \times 100.
\]  

(11)

In the following Tables I–III, the relative gain scores are reported for each of the associated models. In the results related to the time-independent models, such a comparison is not extremely informative. Rather in this first comparison, we assess what is the most indicative baseline model to perform the subsequent comparisons.

B. UPA-Only Time-Independent Models

The performance of time-independent models is reported in Table I. FC-AE performs as well as classical regression models. In a time-independent scenario, where the interest is only the prediction of a wind speed label given a single acoustic spectrum, a framework like the one described in (1) could then be preferable. Since in the time-independent configuration, the FC-AE and the CatBoost have similar performance, see Table I; the performance metric of these models is now taken as the baseline to evaluate the improvement hereafter. In this table, no data are provided for the mean and quartiles for CatBoost and Random Forest since these are ensemble methods, and hence, they give yet an aggregated output.

C. UPA-Only Time-Dependent Case

Table II displays the results given by the model detailed in Section IV. The case of a single-modal data set presents a relative improvement of the performance with respect to the time-independent configurations. FC-AE applied in a time-dependent scenario has a similar performance as in the time-independent case. Indeed, the Conv-AE model yet suffices to improve the time-independent FC-AE baseline by 74%. The 4DVarNet-UPA leads to a relative gain of 8.4% with respect to the FC-AE and regression model baselines.

D. Multimodal Time-Dependent Case

Results in Table II confirm the potential of the multimodal approach. The Conv-AE trained on a heterogeneous data set gives a gain of 12.3% with respect to the FC-AE benchmark. The 4DVarNet-UPA+ECMWF model, the multimodal counterpart of the single-modal 4DVarNet-UPA, yields a performance gain of 14.7% and 15.8% with respect to the FC-AE benchmark and the state-of-the-art presented by the previous work. Recall that such a result by the 4DVarNet was obtained through the training protocol explained in Section IV-C: 5 gradient iterations on the first training step and 10 iterations on the second step after selection by the best score on the validation set. Fig. 3 presents a visual comparison between ground truth wind speed time series and the reconstructions obtained with selected models. ECMWF wind values are also superimposed. These time series are compared with the reconstruction performed by a time-independent and a time-dependent model, the CatBoost, and the multimodal 4DVarNet, respectively.

Fig. 4 shows a scatterplot between the reconstructed and ground truth values. Fig. 4(a) left panel clearly highlights a bias

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Fig. 3. Wind speed reconstructions: Examples of wind speed reconstruction over different 24-h windows in the test data set. We depict the in-situ wind speed and compare the reconstructions issued from the proposed approach to CatBoost reconstruction [15] and ECMWF wind speed. Each panel represents a 24-h time window of the test set. There is no temporal contiguity between each subplot.

Fig. 4. Overall reconstructions and error patterns. (a) Scatterplots of in-situ ground truths against the reconstructions of wind speed obtained with 4DVarNet and ECMWF wind speeds. (b) Hourly averages of the reconstruction error.

in the reconstruction of high wind speed values. This could be due to the saturation of underwater acoustic data for large wind speeds. Fig. 4(b) presents the average hourly error. This plot shows the effect of 1-D convolutional filters since the boundaries of the time observation intervals are not entirely involved by the striding of the filters; hence, there is no coverage backward and forward. One other interesting feature of this plot is the worse predictive performance on the central part of the day. This might be due to the diurnal cycle of winds trend, with the mildest wind speed during the day and strongest winds during the nights. Additionally, winds modeled by ECMWF are also known to suffer from a bias usually ascribable to a misrepresentation of mesoscale convective variability and wind shear [51].

E. Multimodal Time-Dependent Case With Missing Data

One of the points of interest of 4DVarNet is that it can handle time series with missing and/or corrupted data. Since underwater acoustic data in our data set are almost complete, experiments on missing data were designed by arbitrarily artificially masking data batches during training and testing. The missing data percentage is a parameter chosen to range between 10% and 90%. Table III reports the complete set of our experiments. Note that when 10% of the data is missing, 4DVarNet performs as good as the best multimodal model trained with complete data. One may argue that removing 10% of data is analogous to dropout mechanics, artificially removing features, and/or noising the data samples [52], [53]. We may also note that the proposed approach reaches almost the same performance as the CatBoost model presented in [15] up to 70% of missing data.

F. A-Posteriori Classification Performance

We propose a complementary analysis to assess the performance of our model in a classification perspective. For operational applications, it is more practical to express the wind speed as wind classes rather than numerical values. An empirical way to classify wind speed categories is the Beaufort scale, a subdivision based on observable effects of wind on land and sea scenarios. The Beaufort scale is composed of 12 classes. The class 0, called calm is associated with wind speed values lesser than 0.5 m/s. The class 12, called Hurricane-force, is associated with the wind speed greater than 32.7 m/s. Due to the wind speed distribution of the test set, visualized in Fig. 1(c), we choose to define three macroclasses as follows. The first class, “low wind” comprehends the wind speed values ranging from 0 to 3.4 m/s. In terms of the Beaufort scale, these are winds up to the gentle breeze. The second class “median wind” comprehends wind speeds from 3.4 to 10.8 m/s; this latter is the strong breeze category according to the Beaufort scale. The last class “high wind” comprehends all the wind speed values stronger than the strong breeze, larger than 10.8 m/s.

The a-posteriori analysis on wind speed classification shows that the 4DVarNet model trained on regression performs better in
terms of classification accuracy with respect to CatBoost and the multimodal direct inversion model. Fig. 5 reports the confusion matrices of the classification task for each of these three models. The diagonal structure of the matrix is evident for each case. The multimodal 4DVarNet has a larger precision for medium and high wind values. The average accuracy scores for each class are 89%, 91%, and 93% for the CatBoost, multimodal direct inversion, and multi-modal 4DVarNet, respectively. These results show that our framework could be valuable as an instrument to provide reliable qualitative estimates of wind speed, offering operationally useful information about the atmospheric conditions. To conclude, we may remark that a proper classification task on wind speed categories could be approached by designing the model, the data set, and the evaluation criterion in such a way as to account explicitly for the wind speed values in the form of classes. That would be a classification problem rather than a regression problem, as done in this case study.

VI. CONCLUSION

This article presented a novel, robust, and efficient framework for managing the time dependence of acoustic data sets used for the estimation of wind speed values. While previous work successfully demonstrated that machine learning approaches are promising tools to perform wind speed estimation given underwater acoustic data, this work highlights and proves the importance of explicitly accounting for time dependence. This concept could be further expanded to short-term, then to long-term, forecasting problems, that is predicting a time series of wind speed given a small amount of acoustic data, in such a way to forecast the wind trend in a near-future window. We believe that the effective representation of time dependence is particularly relevant, as in an operational scenario, no data are available after a given time. In that case, the model used should be capable of predicting wind speed in the near future. For this task, a complete characterization of wind speed dynamical behavior is necessary.

Further work may also consider the joint use of acoustic data and satellite imagery, such as synthetic aperture radar (SAR) images. While acoustic data have limited spatial coverage but have a rich resolution in time, SAR images display opposite characteristics, as they are scarce in time but offer a wider spatial resolution. A multimodal approach that bridges these two temporally-rich and spatially-rich features could lead to interesting research directions aiming to fit trainable models in which learning one modality helps in learning the other. Cross-modal learning and generation is a salient and important feature of multimodal machine learning [22].

A further improvement could address the target variable to be modeled. In this case study, we mainly focused on the prediction of an environmental variable, but other important applications could benefit from the use of underwater ambient noise for anthropic activities, such as submarine recognition or sea wildlife observation.

REFERENCES

[1] X.-M. Li, “Retrieval of sea surface wind speed by spaceborne SAR based on machine learning,” in Proc. IGARSS IEEE Int. Geosci. Remote Sens. Symp., 2020, pp. 4011–4014.
[2] H. Hersbach, “Comparison of C-band scatterometer CMOD5.N equivalent neutral winds with ECMWF,” J. Atmospheric Ocean. Technol., vol. 27, no. 4, pp. 721–736, 2010.
[3] F. M. Rana, M. Adamo, R. Lucas, and P. Blonda, “Sea surface wind retrieval in coastal areas by means of Sentinel-1 and numerical weather prediction model data,” Remote Sens. Environ., vol. 225, pp. 379–391, 2019.
[4] J.-C. Jang, K.-A. Park, A. A. Mouche, B. Chapron, and J.-H. Lee, “Validation of sea surface wind from Sentinel-1A/B SAR data in the coastal regions of the Korean Peninsula,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 12, no. 7, pp. 2513–2529, Jul. 2019.
[5] C. Wang et al., “A labelled ocean SAR imagery dataset of ten geophysical phenomena from Sentinel-1 wave mode,” Geosci. Data J., vol. 6, no. 2, pp. 105–115, 2019.
[6] M. Zuba, E. Anagnostou, J. Nystuen, and M. Anagnostou, “UPAL: Underwater passive aquatic listener,” in Proc. OCEANS—IEEE Int. Symp., 2015, pp. 1–4.
[7] J. A. Nystuen, M. N. Anagnostou, E. N. Anagnostou, and A. Papadopoulos, “Monitoring Greek seas using passive underwater acoustics,” J. Atmospheric Ocean. Technol., vol. 32, no. 2, pp. 334–349, 2015.
[8] J. A. Nystuen, “Rainfall measurements using underwater ambient noise,” J. Acoustical Soc. Amer., vol. 79, no. 4, pp. 972–982, 1986.
[9] B. B. Ma and J. A. Nystuen, “Detection of rainfall events using underwater passive acoustic sensors and air-sea temperature changes in the tropical Pacific Ocean,” Monthly Weather Rev., vol. 135, no. 10, pp. 3599–3612, 2007.
[10] J. A. Nystuen, S. E. Moore, and P. J. Staben, “A sound budget for the southeastern Bering Sea: Measuring wind, rainfall, shipping, and other sources of underwater sound,” J. Acoustical Soc. Amer., vol. 128, no. 1, pp. 58–65, 2010.
Dorian Cazau received the Ph.D. degree in acoustics and signal processing from Pierre-and-Marie-Curie University, Paris, France, in 2012. In 2015, he held a postdoctoral position with ENSTA Bretagne, Brest, France, and, in 2018, another postdoctoral position with the Institut Mines Telecom Atlantique, Brest. Since 2020, he has been an Assistant Professor with Lab-STICC (UMR CNRS 6285), ENSTA Bretagne. His research revolves around the development of innovative methods of the underwater passive acoustical data, promoted in various communities including marine ecology, bioacoustics, meteorology, oceanography, and seismology. Since 2018, he has also been the PI of the OsmOSE project (https://osmose.ifremer.fr/) aiming to put forward best practices in open sciences and facilitate interdisciplinary research.

Nicolas Farrugia received the electrical engineering degree in signal processing and the M.Sc. degree in signal and image processing from ENSEA, Cergy-Pontoise, France, both in 2005, and the Ph.D. degree in parallel embedded face detection from the Université de Bourgogne and Orange Labs, Grenoble, France, in 2008, working on the hardware implementation of deep neural networks.

In 2010, he moved to the field of cognitive neurosciences as a Postdoctoral Researcher, with a focus on the auditory cognitive neurosciences using methods such as electroencephalography and functional MRI, as well as behavioral psychology methods and motion capture. He joined IMT Atlantique in 2016 as an Associate Professor to engage in a transdisciplinary effort, combining methods from cognitive neurosciences, deep/machine learning, and graph signal processing. His research interests include developing innovative methods to better understand sounds and the brain using modern machine learning and deep learning.

Alexandre Gensse received the digital electronics and computer engineering degree from ISEN Toulon, Toulon, France, in 2014. He was the Scientific Computing Engineer in 2014 and is currently the Data-Scientist for Naval Group. He is currently conducting research for AI applications on naval defense use cases. Most of his work is dedicated to underwater acoustic data processing and classification.

Sara Pensieri received the graduate degree in telecommunication engineering and the Ph.D. degree in electronic and computer engineering, robotics, and telecommunications from the University of Genoa, Genoa, Italy, in 2005 and 2013, respectively. Since 2005, she has been a Researcher with the National Research Council of Italy, the Institute for the Study of the Anthropic Impacts and the Sustainability of the Marine Environment, Genoa. She is involved in the scientific and technological development of the W1M3A observatory. Her expertise spans from operational oceanography to underwater acoustics and the analysis of physical processes at the air-sea interface. She was involved in several European Projects (MERSEA, EuroSITES, MyOCEAN, PERSEUS, FixO2, MINKE) aiming at the development of innovative methodologies for the marine environmental monitoring from unmanned platforms. She has experience in data management and in the design of field experiments at sea during oceanographic cruises.

Roberto Bozzano received the electronic engineer degree in electronic engineering from the University of Genoa, Genoa, July, in 1993. Since 1996, he has been with the National Research Council of Italy, Genoa. His theoretical and experimental research activity concerns the field of complex and multisensory technologies for environmental monitoring with particular reference to the marine-maritime sector. Specifically, the activity concerns the design of complex multifunctional measurement systems and their remote management, the development of acoustic methodologies for the measurement of environmental parameters, the study of the physical processes of energy exchanges at the interface-sea, and, in general, the development of systems for the acquisition and processing of signals on real-time embedded platforms and architectures.

Mr. Bozzano is the Principal Investigator of the offshore observatory W1M3A, part of ERIC-EMSO and ERIC-ICOS. He is also an active member of regional and global alliances, such as MonGOOS and OceanSTITES.

Ronen Fablet (Senior Member, IEEE) received the engineer degree in space and aeronautics from the École Nationale Supérieure de l'Aeronautique et de l’Espace, Toulouse, France, in 1997, and the Ph.D. degree in signal processing and telecommunications from the University of Rennes/INRIA Rennes, Rennes, France, in 2001.

In 2002, he was an INRIA Postdoctoral Fellow with Brown University, Providence, RI, USA. From 2003 to 2007, he held a full-time research position with IFREMER, Brest, in the field of signal and image processing applied to fisheries science. In 2008, he joined the Signal and Communications Department, IMT Atlantique Bretagne-Pays de la Loire (formerly Télécom Bretagne), as an Associate Professor, and has been holding a full Professor position since 2012. He was a Visiting Researcher with Institut de Recherche pour le Développement and IMARPE (Peruvian Sea Research Institute, Peru) in 2011 and a Visiting Professor with IMEDEA (CSIC/UIB, Spain) in 2016. He has led national and international research initiatives at the crossroads of data science, ocean observation, and ocean modeling. His current research interests include physics-informed deep learning for ocean modeling, monitoring, and surveillance.