A Multi-Stage Model for Dissolved Oxygen Monitoring of Coastal Seawater

A multi-stage model for Dissolved Oxygen monitoring of coastal seawater

Vito Ferri
Politecnico di Torino
Torino, Italy
vitoferri95@gmail.com

Sele Okeoghene Thomas
Politecnico di Torino
Torino, Italy
seloke.thomas@gmail.com

Andrea Bordone
Marine Environment Research Centre of S. Teresa ENEA
Pozzuolo di Lerici, Italy
andrea.bordone@enea.it

Giancarlo Raiteri
Marine Environment Research Centre of S. Teresa ENEA
Pozzuolo di Lerici, Italy
giancarlo.raiteri@enea.it

Tiziana Ciuffardi
Marine Environment Research Centre of S. Teresa ENEA
Pozzuolo di Lerici, Italy
tiziana.ciuffardi@enea.it

Chiara Lombardi
Marine Environment Research Centre of S. Teresa ENEA
Pozzuolo di Lerici, Italy
chiara.lombardi@enea.it

Chiara Petrioli
W·S E N S E and Sapienza University Computer Engineering Department
Roma, Italy
chiara.petrioli@uniroma1.it

Daniele Spaccini
W·S E N S E
Roma, Italy
daniele.spaccini@wsense.it

Petrika Gjanci
W·S E N S E
Roma, Italy
petrika.gjanci@wsense.it

Francesca Pennecchi
Ist. Naz. di Ricerca Metrologica
Torino, Italy
f.pennecchi@inrim.it

Marco Coïsson
Ist. Naz. di Ricerca Metrologica
Torino, Italy
m.coisson@inrim.it

Gianfranco Durin
Ist. Naz. di Ricerca Metrologica
Torino, Italy
g.durin@inrim.it

***Abstract*—We propose a multi-stage model for monitoring the Dissolved Oxygen measured continuously (every half an hour) by underwater sensors in the "Smart Bay Santa Teresa", located on the Ligurian Eastern coast near La Spezia. This model represents the first attempt to construct a local digital twin of the bay, and it is based on three separated models for Water Temperature, Pressure (Depth) and Conductivity (Salinity). This approach allows us to reconstruct missing values of the Dissolved Oxygen in case of problems and failures, and also to correct the effect of biofouling on the sensors. Our procedure aims at establishing a flexible framework that can be applied across various coastal environments, by leveraging both underwater sensor data and meteorological information to generate accurate descriptions and future predictions tailored to the specific study area.**

***Index Terms*—Remote sensing, Time Series Forecasting, Neural Networks, Machine Learning, Dissolved Oxygen, Seawater, Data Cleansing, Biofouling**

I. Introduction

During the One Ocean meeting in Brest in 2022 [1], the President of the European Commission von der Leyen announced three new key initiatives to restore and preserve the oceans: a new international coalition to protect biodiversity on high seas; the EU's research mission to restore our ocean and waters by 2030; and finally an ambitious project to digitally simulate the world's oceans, known as "European Digital Twin Ocean" (EDTO). This is a plan to make available to scientists, entrepreneurs and society in general all the ocean data, advanced analytical models and simulation tools, as "an interactive replica of the ocean for a better decision-making".

Marine data and advanced models are the key ingredients for the success of this initiative. Artificial Intelligence (AI) models in particular are progressively used by researchers to investigate the connections between the (essential) ocean variables, in order to be able to make reasonable predictions and simulate possible scenarios, and take the best informed decisions [2]. This is extremely important for the growth of a blue economy, to face and mitigate climate changes, and to protect coastal communities and habitats.

Based on existing services like Copernicus Marine Service [3], and the European Marine Observation and Data Network [4], the EDTO will act on different time and space scales, with generic ocean models and AI toolboxes on top of "local digital twins", to better adapt to small scales events and peculiarities. In this paper, we aim to present the first step to build a local digital twin of the "Smart Bay Santa Teresa", a bay located on the Ligurian Eastern coast near La Spezia (see Fig. 1). Our focus here is on the analysis and modelling of Dissolved Oxygen detected using the Internet of Underwater Things (IoUT) technology [5], which allows us to get values of the Oxygen almost continuously (every half an hour), together with other physical quantities, such as water temperature, salinity, etc.

Monitoring Dissolved Oxygen levels in coastal regions is particularly significant for determining seawater quality [6], the marine life activity [7], [8], and for aquaculture [9], [10]. Due to global warming, oxygen concentrations are declining in both open ocean and coastal waters since the middle of the

TABLE I
AVAILABLE SENSORS IN THE SMART BAY AREA WITH SAMPLING RATES

SENSOR	MODEL	PARAMETERS	SAMPLING
IoUT sensors	AANDERAA 4319	Water Pressure, Temperature, Conductivity	30 minutes
	AANDERAA Optode 4531	Dissolved Oxygen	
CTD probe	SeaBird SBE19plus	Water Pressure, Temperature, Conductivity	7 Days (approx.)
	SeaBird SBE43	Dissolved Oxygen	
	Turner Designs Cyclops-7 CHL / TUR	Fluorescence / Turbidity	
Meteorological station	Lastem C110R / C100A	Solar radiation / Precipitation	10 Minutes
	Vaisala PTB101B	Atmospheric Pressure	
	Rotronic MP101A	Air Temperature, Relative Humidity	
	Didcot DWD-103 / DWR-201G	Wind Direction / Speed	

20th century, so it is extremely important to forecast long-term changes especially in terms of deoxygenation, and prevent hypoxia, in order to forecast future crisis and properly protect the environment.

Continuous monitoring can be difficult, due to various factors, such as loss of signals and connections with discontinuities in transmission of temporal data, electrical anomalies, and noisy signals of different origin [11]. In addition, biofouling and other bio-factors can alter significantly the performance of the sensors. All these complications have the potential to compromise the reliability of the collected data, thereby impacting the accuracy of any subsequent analysis. Here we show an advanced data cleansing strategy based on AI modelling to filter existing data and recover the missing ones. Our strategy is based on a multi-stage approach, which makes use of separated models for Water Temperature, Pressure (Depth) and Conductivity (Salinity), respectively, to feed a final model able to describe with accuracy the variations of Dissolved Oxygen in the bay.

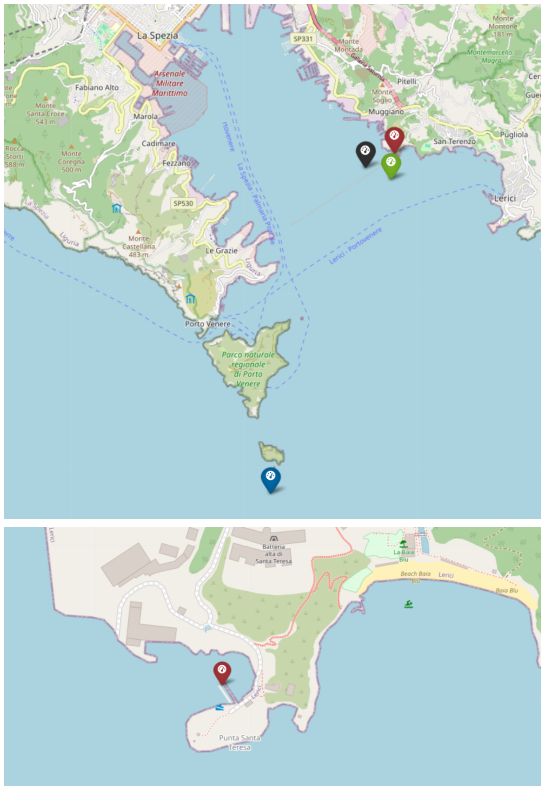

Fig. 1. The "Smart Bay Santa Teresa" located on the Eastern Ligurian coast, near La Spezia. (top) Coloured icons show the different nodes of measurements: the blue node close the Tinetto island at a depth of 16 m, the black and green nodes across the land bridge (the slim white line) at 5 m, and the red main node at 1 m, close to the coast. (bottom) The main node used in this study is located within the bay and under the jetty. (Map done using Folium and OpenStreetMaps [12])

II. MARINE AND ATMOSPHERIC DATA

A. Data collection

The data were collected as part of a collaborative project involving WSense (an Italian technology company specializing in underwater monitoring and communication systems), ENEA (Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile, the Italian National Agency for New Technology, Energy, and the Environment), and INRiM (Istituto Nazionale di Ricerca Metrologica, the Italian National Institute of Metrological Research). High-temporal resolution data (at half an hour rate) were collected using IoUT sensor networks from WSense at the main node of Fig. 1, while weekly data come from a multiparametric conductivity-temperature-depth (CTD) probe calibrated at the ENEA laboratories and deployed in the sea close to the WSense main node In addition, a ENEA-operated Meteorological station, located at an altitude of 50 m in the same area, collects solar radiation, atmospheric pressure, air temperature, wind direction and speed, and relative humidity every 10 minutes (see Tab. I for instrument models). Data acquisition spans 21 months, from December 2021 to August 2023, when the sensors were removed from the bay. As little differences in the time acquisition can occur, we set a common temporal

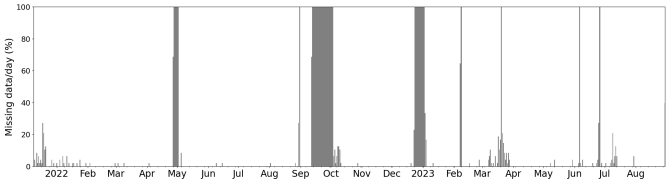

Fig. 2. Occurrence of missing data during the full period of 21 months of measurements with the IoUT Sensors. The vertical bars indicate the percentage of missing data for each day.

base by resampling all the data hourly by averaging, with the exception of solar radiation and rain which are summed within each hour. This reduces the noise in the data and the distortion of the outliers, while preserving the daily components of the data.

B. Data anomalies and noise

Different anomalies affect the acquisition and the quality of the data in such a harsh environment: i) Missing data, mostly due to the maintenance of nodes, while a small percentage, around 4-5% of total data generated by the nodes, due to electrical failures, lack of underwater connectivity, or sensor failures. Fig. 2 shows the percentage of Oxygen values missing during the 21 months of acquisition, where three large periods can be clearly identified, rendering the data unsuitable for use in Machine Learning (ML) / Deep Learning (DL) models that require continuous time series data. While small gaps can be easily addressed through interpolation, these large gaps pose a more complex problem that cannot be resolved through simple methods: ii) Biofouling, particularly significant during the bloom season (see Fig. 3). During this period, the light radiation highly affects photosynthetic processes driven by organism in the water column (phytoplankton) and growing on the substrate (i.e., bacteria and macrobenthic photosynthetic organisms). In practice, the amplitude of the daily variations in Oxygen levels due to the photosynthetic activities increase over time. For these reasons, a periodical cleaning of the sensors is needed, which coincides with the weekly measurements with the CTD (purple bullets in Fig. 3 at the cleaning days shown by the vertical dash lines); iii) any other less predictable events, such as a plastic bags covering the sensors' head, can also significantly impact the quality of the data. In addition, sensor under-powering can lead to inaccurate or flawed readings, particularly when operating in challenging aquatic environments where maintenance might be limited or unfeasible.

Meteorological variables are also naturally noisy, due to the normal fluctuations during the evolution of the weather. This behaviour has a negative impact on the performance of the machine learning models. On the other hand, their effects on the seawater properties are generally delayed on time. We found that a reasonable solution is to apply a rolling window technique to the raw data, smoothing out noise while keeping underlying trends over specific periods.

C. Data pre-processing

To mitigate these complications, we implement a stringent quality control protocol. Specifically, data points recorded immediately before and after each cleaning event are excluded from analysis, as these readings may not accurately reflect the true state of the system. At the same time, to mitigate the effect of the biofouling, we exclude 40% of data points closest to the cleaning events if their value significantly deviated from the overall trend observed before and after the cleaning period. In practice, data acquired far from the bloom season may all be included in the analysis, whereas data collected during the bloom, like in lower part of Fig. 3, are highly filtered.

III. The Dissolved Oxygen model

A. The multi-stage approach

Although the WSense network delivered data consistently (above 94% of the generated information from the sensors during the period January–February), building a tuned Dissolved Oxygen model (DOM) considering sporadic missing data due to internal sensor failures and maintenance, turned out to be ineffective preventing the model to converge. Instead, we used a different approach, based on a *multi-stage* cascade procedure where we first build separated models for Water Temperature, Pressure/Depth, and Conductivity/Salinity and then feed the final DOM (see Fig. 4). For each model, we use Ensemble modelling, specifically the Stacking Regressor method [13], an architecture which combines two component models: the Adaptive Boosting (AdaBoost) and Extremely Randomized Trees (ExtraTrees). As the name suggests, stacking involves training a final linear regressor on the combined predictions provided by the underlying base estimators. This mitigates overfitting, and improves the resistance to outliers.

In practice, we start with the Water Temperature model, using the WSense Temperature data as target, and Meteorological and CTD Temperature data as features. As CTD is sampled weekly, we artificially interpolate it at the WSense sampling. In general, the use of CTD data as a feature and not as a target, highly improves the regression especially during the large periods of missing data of Fig. 2. After the Water Temperature, we then build the Pressure and Conductivity models, using the same scheme, i.e. we train the WSense measurements with CTD and Meteorological data as features. Finally, the Oxygen model is built by using the outputs of the three previous models targeting the WSense Dissolved Oxygen data.

As for Valera et al. [14], we highlighted the potential of decision tree-based approaches [15], such as Random Forest [16], to accurately reconstruct and predict Dissolved Oxygen concentrations. There are significant differences, by the way. In our case, we use a combination of low-cost (WSense) sensors for continuous measurements, and high cost CTD sensors for calibration of the models, to estimate the missing data and correct anomalies by using a sequence of models for each variable. The reason for adopting this procedure has the advantage to disentangle the physical from the the

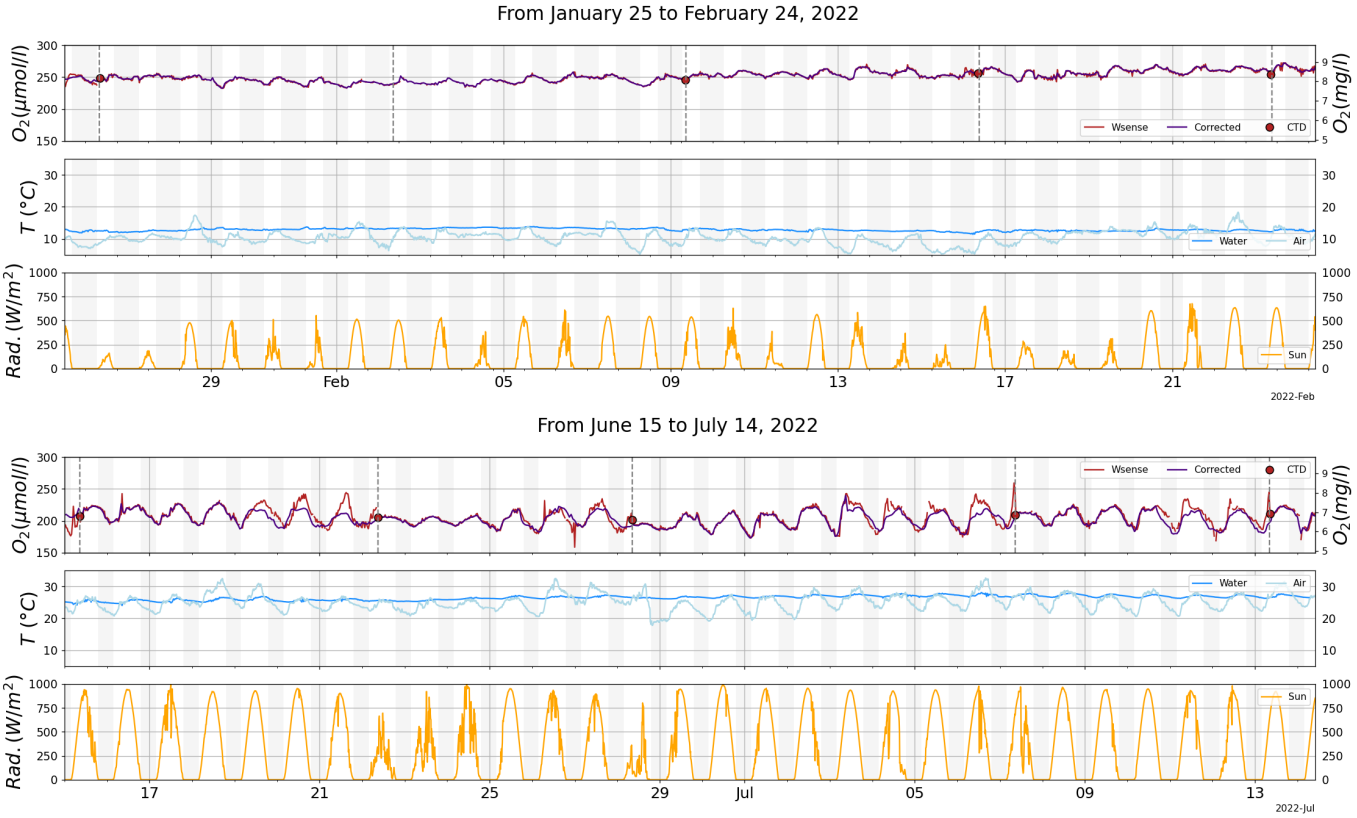

Fig. 3. Examples of measurements of Dissolved Oxygen, Water and Air Temperatures and Solar Radiation from the node close to the coast (red node in Fig. 1, 1 m depth), in winter (upper part), and during the bloom season (lower part). In the latter, the amplitude of the signal grows significantly due to biofouling after a few days from cleaning (dashed vertical lines), while significant drifts can also occur due to the extraction of the sensor from water. Other anomalies in the data are clearly visible: missing points, spikes, noise, etc. In both seasons, the Oxygen increases during the hours of light and decreases roughly linearly on nights (shaded vertical bars).

biological effects: in fact, from the simple physical point of view, the Dissolved Oxygen depends on three variables, i.e. the Water temperature, the Pressure (Depth), and the Conductivity (Salinity). Once they are estimated by the Stacking regressors, we can calculate the thermodynamic Dissolved Oxygen, using the Thermodynamic Equation Of Seawater (TEOS-10) [17], i.e. the value expected in the absence of any biological effect. Any difference with the detected Oxygen signal can be thus ascribed to the phytoplankton production, as seen in Fig. 5 by the black lines. This is particularly significant during the bloom, as expected.

TABLE II
PERFORMANCE METRICS OF THE MULTI-STAGE MODEL

Metric	**Temperature** (°C)	**Pressure** (db)	**Conductivity** (mS/cm)	**Oxygen** (μmol/l)
MSE	0.013	0.002	0.021	16.49
RMSE	0.114	0.012	0.145	4.061
R^2	0.999	0.982	0.999	0.969
CV	0.998	0.959	0.998	0.919

B. Model metrics

A comprehensive overview of the performance metrics for the four regressive models is reported in Tab. II. To evaluate the accuracy and robustness of the proposed regression models, we used the mean squared error (MSE) and its variant, the root mean squared error (RMSE), which quantified the discrepancy between observed and forecast values, with larger errors receiving greater weights. Additionally, the coefficient of determination (R^2), a widely used metric, measured the proportion of the total data variability that was successfully captured by the model. To further validate the models, cross-validation (CV) techniques were implemented, resulting in a robust estimation of the prediction error and enabling the detection of potential overfitting. Our findings indicate that the developed models exhibited very good predictive capabilities, as evident from the evaluated metrics.

IV. MODELLING MARINE ACTIVITY

A significant application of these predictive models concerns the modelling of marine activity, which can make a fundamental contribution to understanding the biological and physical processes taking place in the seawater.

As seen in lower part of Fig. 5, i.e. the differences between theoretical Oxygen values, derived from the model, and those

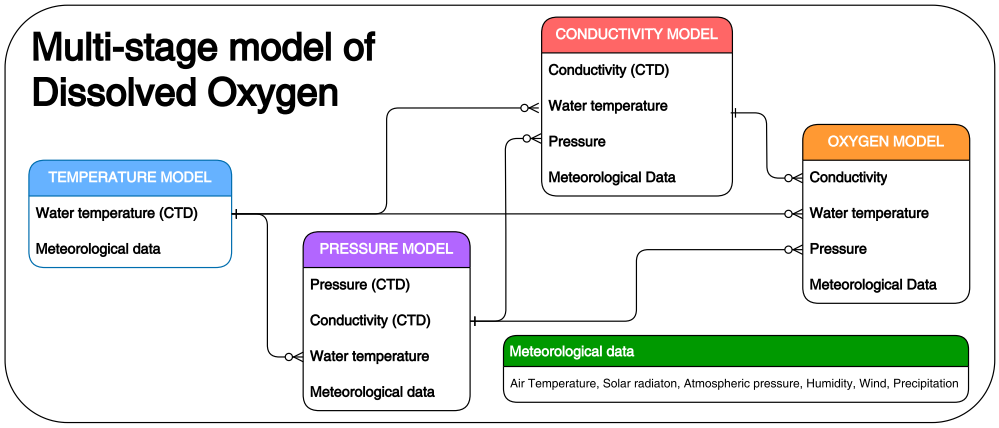

Fig. 4. Schematic diagram of the multi-stage model used to calculate the Dissolved Oxygen concentration using previous trained regression models for Water Temperature, Pressure, and Conductivity, together with meteorological data.

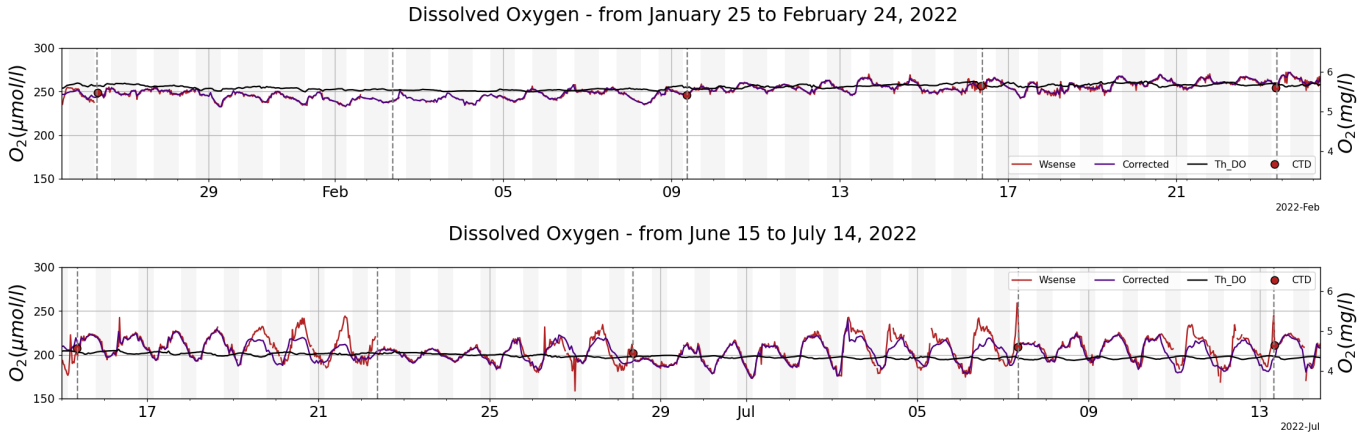

Fig. 5. Examples of the anomalies corrected using the multi-stage regressive model of Dissolved Oxygen for the off-bloom (upper part) and bloom seasons (lower part), as in Fig. 3. Red lines are the WSense data, the blue curves the output of the Oxygen model, while the black lines are the theoretical Dissolved Oxygen (Th_DO), calculated using Thermodynamic Equation Of Seawater (TEOS-10), using the values of Water Temperature, Pressure, and Conductivity [17]. Clearly, the corrections are more significant during the bloom season.

actually measured, can prove to be a valuable indicator of marine activity and its impact on Oxygen levels by isolating the ecosystem-influenced component. This is because Oxygen plays a crucial role in marine ecosystems, influencing both biological and chemical processes. The analysis of these residues can therefore offer interesting insights into the impact of the marine biological component on the production and consumption of Dissolved Oxygen in seawater over time, allowing the identification of any anomalies or emerging trends.

V. Conclusions and Perspectives

This study demonstrates the efficacy of AI models in cleaning and modelling environmental parameters, as we were able to manage and correct anomalies, errors, and irregularities in data sampling. We have shown as various data cleaning operations significantly improved the accuracy of regressions, such as the use of techniques like Rolling Window and error removal through tailored algorithms. In addition, we were able to split the contribution of the biological and physical processes. This clearly demands further investigation. We are also currently investigating the use of these models to forecast the Oxygen level, once we have a reasonable prediction of the Meteorological data. At the moment the model that gives the most promising results is based on Recurrent Neural Networks, in particular the Long Short-Term Memory. To fully realise a local digital twin of the Santa Teresa bay, we clearly need to expand this one-site model to more measurement nodes, at different distances from the coast and different depths. In these very days, an installation of six new nodes in the bay is taking place, that will be trained as soon they are operative.

References

[1] "One Ocean Summit 2022," https://oneplanetsummit.fr/en/events-16/one-ocean-summit-221, accessed: 2024-06-12.

[2] R. Logares, J. Alós, I. Catalán, A. Solana, J. del Campo, G. Ercilla, R. Fablet, A. Fernandez-Guerra, M. Galí Tápias, J. Gasol, A. González, E. Hernández– García, C. López, R. Massana, L. Montiel, M. Palmer, A. Pascual, S. Pascual, F. Perez, and A. Villasenor, *Oceans of big data*

and artificial intelligence, ser. CSIC scientific challenges towards 2030, 01 2021, pp. 163–179.
[3] "Copernicus Marine Service," https://marine.copernicus.eu/, accessed: 2024-06-12.
[4] "European Marine Observation and Data Network (EMODnet)," https://emodnet.ec.europa.eu/, accessed: 2024-06-12.
[5] L. Parra, G. Lloret, J. Lloret, and M. Rodilla, "Physical sensors for precision aquaculture: A review," *Ieee Sens. J.*, vol. 18, no. 10, pp. 3915–3923, may 2018.
[6] M. Burke, J. Grant, R. Filgueira, and T. Stone, "Oceanographic processes control dissolved oxygen variability at a commercial atlantic salmon farm: Application of a real-time sensor network," *Aquaculture*, vol. 533, p. 736143, feb 2021.
[7] V. Di Biagio, R. Martellucci, M. Menna, A. Teruzzi, C. Amadio, E. Mauri, and G. Cossarini, "Dissolved oxygen as an indicator of multiple drivers of the marine ecosystem: the southern Adriatic Sea case study," *State of the Planet*, vol. 1-osr7, p. 10, 2023.
[8] R. K. Walter, S. A. Huie, J. C. P. Abraham, A. Pasulka, K. A. Davis, T. P. Connolly, P. L. Mazzini, and I. Robbins, "Seasonal controls on nearshore dissolved oxygen variability and hypoxia in a coastal embayment," *Estuarine, Coastal and Shelf Science*, vol. 278, p. 108123, Nov. 2022.
[9] A. Chatziantoniou, S. C. Spondylidis, O. Stavrakidis-Zachou, N. Papandroulakis, and K. Topouzelis, "Dissolved oxygen estimation in aquaculture sites using remote sensing and machine learning," *Remote Sensing Applications: Society and Environment*, vol. 28, p. 100865, nov 2022.
[10] X. Ta and Y. Wei, "Research on a dissolved oxygen prediction method for recirculating aquaculture systems based on a convolution neural network," *Comput. Electron. Agr.*, vol. 145, pp. 302–310, Feb. 2018.
[11] M. Hosseinzadeh, E. Azhir, O. H. Ahmed, M. Y. Ghafour, S. H. Ahmed, A. M. Rahmani, and B. Vo, "Data cleansing mechanisms and approaches for big data analytics: a systematic study," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 1, pp. 99–111, 2023. [Online]. Available: https://doi.org/10.1007/s12652-021-03590-2
[12] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org," https://www.openstreetmap.org, 2017.
[13] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, no. 1, pp. 49–64, Jul. 1996.
[14] M. Valera, R. K. Walter, B. A. Bailey, and J. E. Castillo, "Machine learning based predictions of dissolved oxygen in a small coastal embayment," *Journal of Marine Science and Engineering*, vol. 8, no. 12, p. 1007, December 2020.
[15] S. B. Kotsiantis, "Decision trees: a recent overview," *Artificial Intelligence Review*, vol. 39, pp. 261–283, April 2013.
[16] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[17] I. O. C. (IOC), S. C. on Oceanic Research (SCOR), and I. A. for the Physical Sciences of the Oceans (IAPSO), *The international thermodynamic equation of seawater - 2010: Calculation and use of thermodynamic properties*, UNESCO, 2010, manual and Guides No. 56, available from: http://www.TEOS-10.org.