Creating a High-Resolution 3D Oceanographic Digital Twin from Sparse Glider Data using Machine Learning

José Manuel Echevarría Rubio[*]
jechevarriar17@gmail.com

June 15, 2025

Abstract

Abstract: Autonomous underwater gliders are powerful platforms for ocean observation, yet their linear transects result in spatially sparse datasets. This study presents a robust machine learning workflow to transform such sparse data into a complete, high-resolution 3D model—a "digital twin"—of a coastal environment. Using data from a Slocum G2 glider deployed in Trinity Bay, Newfoundland, Canada, between October and November 2014, we developed predictive models for key oceanographic variables, including temperature, salinity, density, and oxygen concentration. Two powerful ensemble algorithms, Random Forest and XGBoost, were trained and validated using the glider's onboard sensor data (GPCTD, Aanderaa Optode) as ground truth. The models were trained to predict variable distribution based on spatial coordinates (latitude, longitude, depth). The resulting predictions were constrained to the physical boundaries of the bay by integrating high-resolution bathymetry from the ETOPO1 dataset and by applying a DBSCAN clustering algorithm to mask the data to the main basin. The models demonstrated exceptional performance, with R^2 values exceeding 0.96 for all primary variables. The final output is a high-resolution, physically constrained 3D data volume that allows for exploration of the oceanographic conditions within the bay, providing a powerful tool for scientific analysis, monitoring, and future mission planning.

Contents

[*]NF-POGO Centre of Excellence, Dalhousie University.

1 Introduction

Coastal ocean environments are among the most dynamic and complex ecosystems on Earth, playing a critical role in global climate, biodiversity, and economy. Understanding their three-dimensional structure is paramount for effective management and research. Autonomous underwater vehicles (AUVs), particularly ocean gliders, have revolutionized our ability to collect in-situ data over long durations and large spatial scales [4].

However, a fundamental challenge remains: while gliders provide high-resolution data along their path, these paths are inherently sparse, representing mere lines through a vast volumetric space. This sparsity makes it difficult to fully characterize the oceanographic state of an entire region, such as a bay or shelf sea, at a specific moment in time.

This study addresses this challenge by proposing and demonstrating a machine learning framework to create a ”digital twin” of a coastal bay. We define a digital twin as a high-resolution, three-dimensional data volume that accurately represents the physical state of the environment. Our approach leverages the rich, albeit sparse, data from a glider survey and uses powerful machine learning models to intelligently interpolate this information throughout the entire bay basin.

The primary objective of this work is to develop and validate a method that can:

1. Predict multiple oceanographic variables (temperature, salinity, density, etc.) in 3D space based on glider data.

2. Constrain these predictions to the realistic physical boundaries of the study area, defined by high-resolution bathymetry and coastline data.

3. Produce a final, explorable 3D data product that can serve as a powerful tool for scientific visualization and analysis.

By transforming a linear dataset into a full volumetric model, this methodology unlocks the true potential of glider surveys, enabling a more holistic understanding of complex coastal systems.

2 Data and Methods

2.1 Study Area

The study was conducted in Trinity Bay, a large, deep bay located on the northeast coast of the island of Newfoundland, Canada (Figure 1). The bay is a significant feature of the region’s coastal oceanography, influenced by the cold Labrador Current and local freshwater runoff, creating complex and dynamic hydrographic conditions.

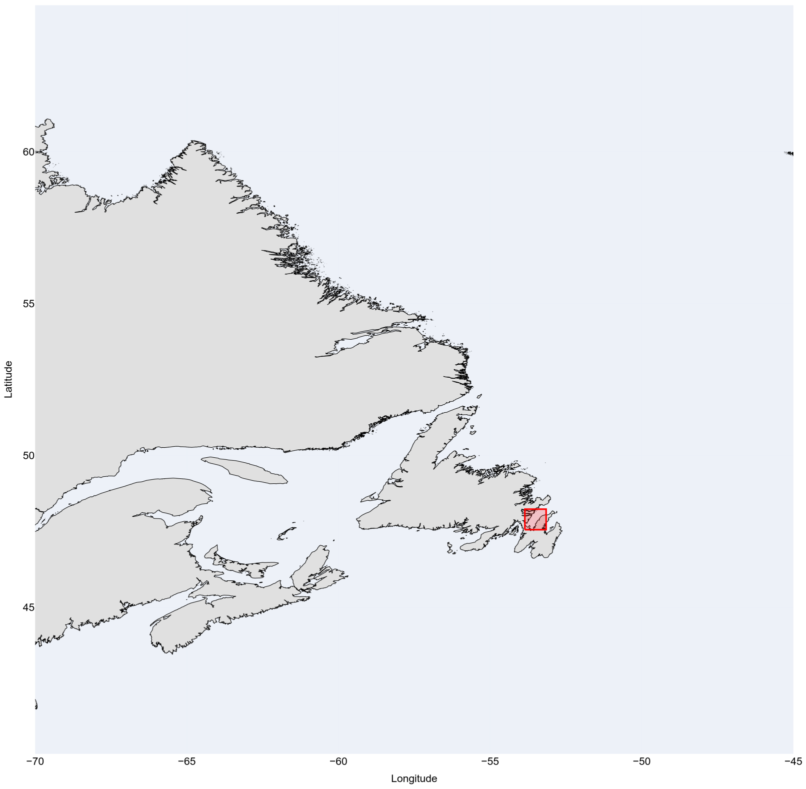

Figure 1: Study area. The map highlights the northeast coast of Newfoundland, with the red box indicating the specific domain of the Trinity Bay glider survey.

2.2 Glider Data Acquisition

The primary dataset was collected by a Teledyne Webb Research Slocum G2 1000m ocean glider, 'Unit 473', operated by Memorial University (MUN). The three-dimensional path of the glider is visualized over the regional bathymetry in Figure 2.

- **Deployment Period:** The survey took place between 01 October 2014 and 05 November 2014.
- **Onboard Sensors:** The glider was equipped with a standard sensor suite, including a pumped Sea-Bird Scientific Glider Payload CTD (GPCTD) for measuring conductivity, temperature, and depth (pressure), and an Aanderaa Optode 4831 for measuring dissolved oxygen concentration.
- **Data Access:** The NetCDF data used in this study was made publicly available by Memorial University through the Canadian Integrated Ocean Observing System (CIOOS) platform [1]. The specific dataset used covers the period from October 28 to November 4, 2014.

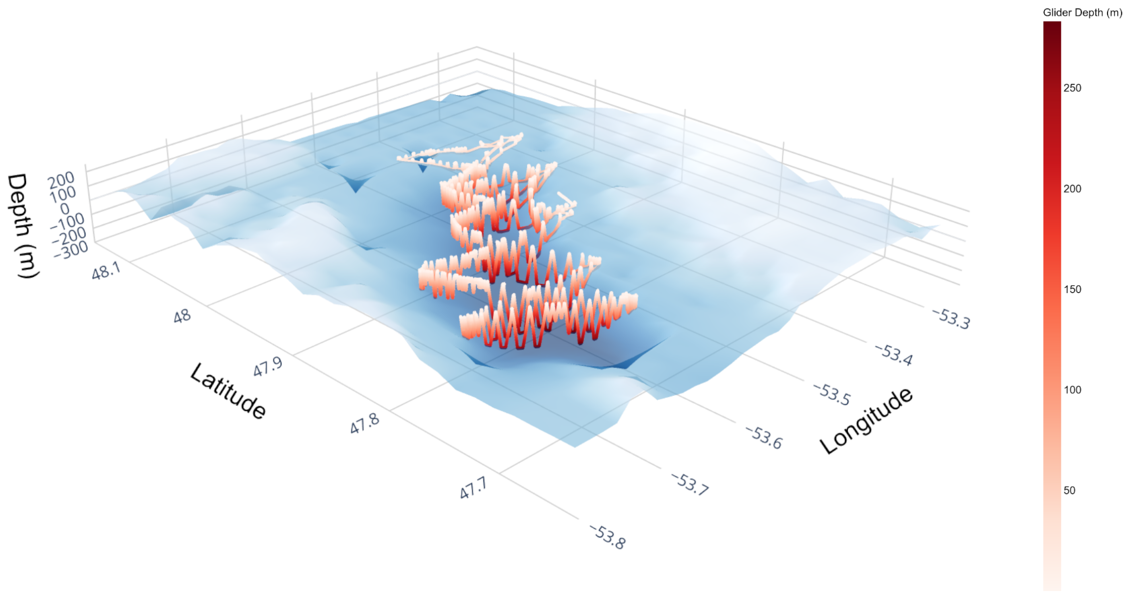

Figure 2: The 3D trajectory of the Slocum glider survey within Trinity Bay (15X vertical exaggeration). The path is colored by depth (red being deeper) and is constrained by the ETOPO1 underlying surface.

2.3 Ancillary Data

To provide geographic and physical context, two ancillary datasets were used:

- **Bathymetry:** Seafloor depth data was obtained from the ETOPO1 Global Relief Model, a 1 arc-minute resolution grid produced by the National Oceanic and Atmospheric Administration (NOAA) [2]. This was used to define the lower boundary of our 3D model.

- **Coastline:** High-resolution coastline data was sourced from the Global Self-consistent, Hierarchical, High-resolution Shoreline (GSHHS) database, a project developed by Wessel and Smith [3]. This was used to geographically mask the data to the bay basin and for visualization.

2.4 Machine Learning Workflow

A comprehensive machine learning pipeline was developed to construct the digital twin.

2.4.1 Model Selection and Features

The task of interpolating data in 3D space is treated as a regression problem. Based on their proven effectiveness with complex, non-linear spatial data, two ensemble learning algorithms were chosen for comparison: **Random Forest Regressor** and **XGBoost Regressor**. For all models, the input features were the spatial coordinates of a given point: `latitude`, `longitude`, and `depth`. A chained prediction approach was used for derived variables.

2.4.2 Model Training and Validation

A rigorous training and validation process was employed for each target variable:

1. **Data Cleaning:** The dataset was filtered to remove any rows containing null values for the required input features and the specific target variable being modeled.

2. **Logarithmic Transformation:** For variables that are physically constrained to be non-negative, such as ‘oxygen_concentration‘, a logarithmic transformation (‘log(1+x)‘) was applied to the target data before training. This ensures the model cannot produce physically impossible negative predictions after the inverse transformation is applied to its outputs.

3. **Hyperparameter Tuning:** A ‘GridSearchCV‘ with 3-fold cross-validation was used to find the optimal hyperparameters for both the Random Forest and XGBoost models.

4. **Model Selection:** The final model for each variable was selected by comparing the out-of-sample Root Mean Squared Error (RMSE) of the tuned models. The model with the lower RMSE was chosen as the best predictor for that variable.

2.4.3 Digital Twin Generation

The final 3D data volume was constructed through a multi-step process:

1. **Grid Creation:** A high-resolution 3D grid of points was generated, spanning the latitude, longitude, and depth ranges of the glider survey.

2. **Bathymetric Filtering:** The ETOPO1 bathymetry data was used to create a seafloor depth interpolator. All points on the grid that fell below the interpolated seafloor were discarded.

3. **Geographic Masking:** The remaining points were then clustered based on their 2D coordinates using the DBSCAN algorithm. The largest contiguous cluster, representing the main basin of Trinity Bay, was identified, and all other points were discarded.

4. **Prediction:** The final set of trained models was used to predict the value of each oceanographic variable at every point within this final, clean ”geobody”.

5. **Data Export:** The resulting high-resolution 3D data volume was exported as a NetCDF file.

3 Results

3.1 Model Performance

The machine learning models demonstrated exceptional predictive power for most variables. XGBoost was selected as the superior model for temperature, salinity, density, and oxygen concentration, while Random Forest performed better for pressure, conductivity, and oxygen saturation. Table 1 summarizes the performance metrics for the best-performing algorithm selected for each variable.

Table 1: Performance metrics of the best-selected models on the test dataset. The R^2 score indicates the proportion of the variance in the target variable that is predictable from the spatial coordinates.

Variable	Best Model	RMSE	R^2 Score
Temperature	XGBoost	0.2945	0.9944
Salinity	XGBoost	0.0313	0.9992
Pressure	Random Forest	0.8059	1.0000
Oxygen Concentration	XGBoost	–	0.9628
Conductivity	Random Forest	0.0005	1.0000
Density	XGBoost	0.0082	1.0000
Oxygen Saturation	Random Forest	0.0623	1.0000

Note: RMSE for Oxygen Concentration is not shown as it was evaluated in log-transformed space and is not directly comparable to the other variables. Model selection was based on performance in that space, with the R^2 score remaining a valid comparative metric.

The extremely high R^2 scores indicate a strong spatial auto-correlation, which is well-captured by the models. The performance is visually confirmed in Figure 3, which shows a tight clustering of predicted temperature values along the 1:1 line with the actual measured values.

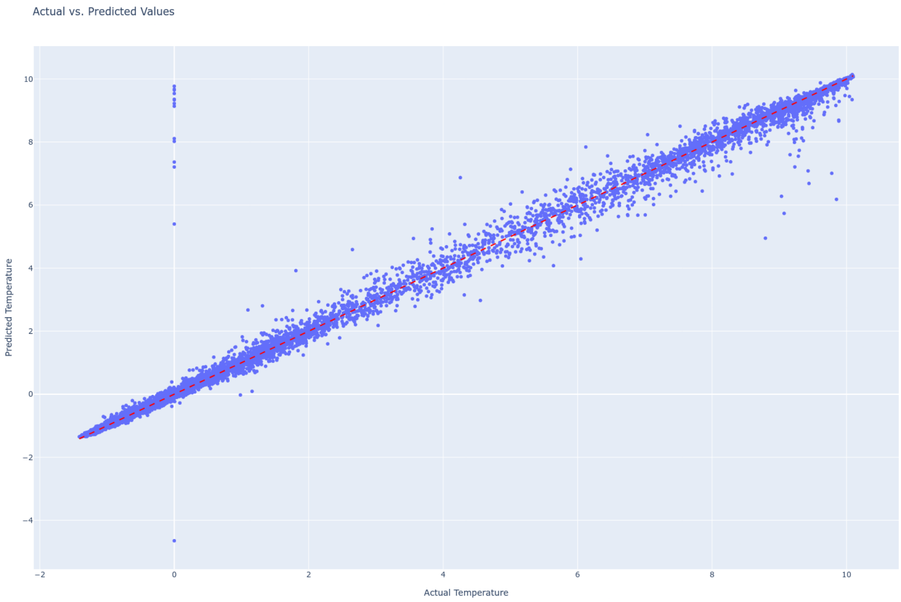

Figure 3: Actual vs. Predicted values for the ‘temperature‘ model on the held-out test set. The tight clustering around the red 1:1 line indicates very high model accuracy.

3.2 The 3D Digital Twin

The final output of the workflow is a high-resolution ”geobody” that is constrained by the real-world bathymetry. Figure 4 provides a powerful visualization of the complete digital twin output. This volumetric dataset can be explored through interactive 3D plots (Figure 5) or by generating cross-sectional animations.

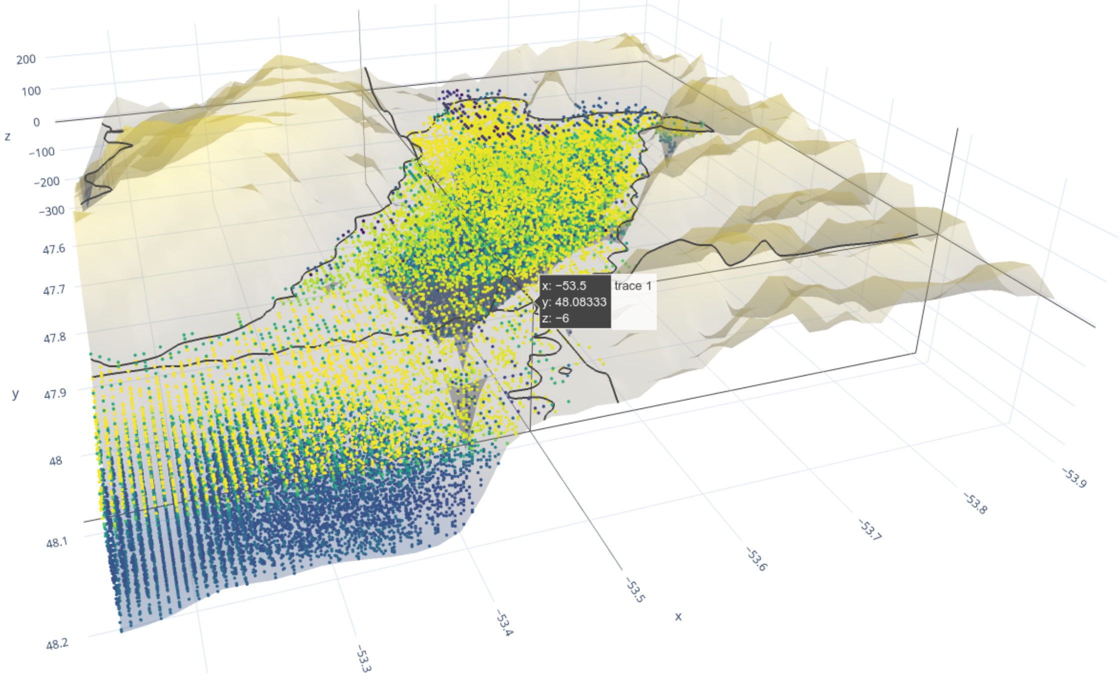

Figure 4: The complete 3D digital twin of predicted seawater temperature, filling the entire bay basin.

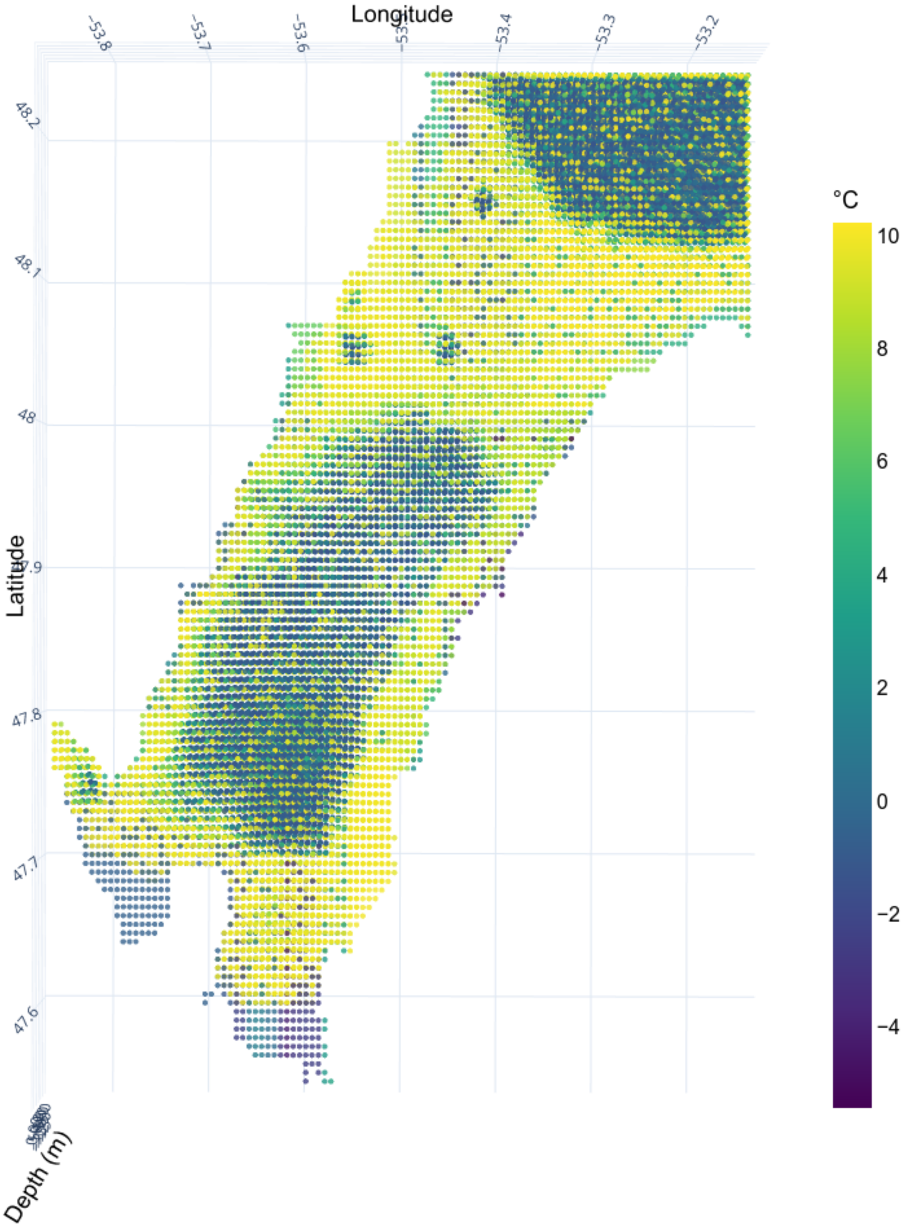

Figure 5: A static visualization of the final predicted temperature geobody. The data volume is constrained by the seafloor bathymetry and the high-resolution coastline, providing a complete and physically realistic model for exploration.

4 Discussion and Applications

The results demonstrate that a machine learning workflow can successfully and accurately transform a sparse glider transect into a complete 3D oceanographic model. The high performance of the models suggests a strong, stable spatial structure in Trinity Bay during the survey period, which is well-captured by the regression models.

The primary application of this digital twin is the ability to perform **virtual sampling**. Researchers can now query the model for an estimated profile of temperature, salinity, or density at any location within the bay, not just where the glider traveled.

This is invaluable for:

- **Feature Identification:** Locating and characterizing the spatial extent of features like the thermocline, halocline, or subsurface chlorophyll maxima.
- **Habitat Modeling:** Providing complete environmental data as input for models that predict the distribution of marine species.
- **Data Validation:** Comparing the model's predictions to data from other assets (e.g., moorings or satellite data) to assess agreement.
- **Future Mission Planning:** Using the 3D model to identify areas of high variability or scientific interest to target with future glider deployments.

It is important to acknowledge the limitations of this approach. The model represents a quasi-synoptic snapshot of the bay; it does not capture temporal evolution over hours or days. Furthermore, the model's accuracy is highest within the domain of the glider's sampling. Predictions in areas far from the glider's path, or in very shallow, unsampled waters, should be treated with greater caution. The poor predictive performance for current velocities ('u', 'v'), which were skipped due to low data volume, highlights that some dynamic variables may require time-series models or different input features beyond simple spatial coordinates.

5 Conclusion

This study has successfully demonstrated a robust, end-to-end workflow for creating a high-resolution 3D digital twin of a coastal bay from sparse ocean glider data. By integrating machine learning with high-resolution bathymetric and coastline data, we produced a physically constrained and geographically accurate model of key oceanographic variables. The resulting data product overcomes the inherent spatial limitations of glider surveys and provides a powerful new tool for comprehensive scientific analysis, visualization, and monitoring of complex coastal ocean environments.

Acknowledgments

We acknowledge Memorial University of Newfoundland for providing the glider data, made accessible through the Canadian Integrated Ocean Observing System (CIOOS). We also thank the creators and maintainers of the ETOPO1 (NOAA) and GSHHS (Wessel and Smith) datasets for making these critical resources publicly available.

References

[1] Canadian Integrated Ocean Observing System (CIOOS). (2021). *MUN Glider Deployment: Unit 473, Trinity Bay, 2014-10-01.* CIOOS Atlantic. Dataset. `https://doi.org/10.17882/79349`

[2] Amante, C. and Eakins, B. W. (2009). *ETOPO1 1 Arc-Minute Global Relief Model: Procedures, Data Sources and Analysis.* NOAA Technical Memorandum NESDIS NGDC-24. `https://doi.org/10.7289/V5C8276M`

[3] Wessel, P., and Smith, W. H. F. (1996). A global, self-consistent, hierarchical, high-resolution shoreline database. *Journal of Geophysical Research*, 101(B4), 8741-8743. https://doi.org/10.1029/96JB00104

[4] Rudnick, D. L., et al. (2004). Underwater gliders for ocean research. *Marine Technology Society Journal*, 38(1), 73-84. https://doi.org/10.4031/002533204787522703