Anomalies In The Depths: Evaluating Unsupervised Anomaly Detection on Real-World Unlabelled Data. A Case Study on Autonomous Ocean Gliders

Federico Scarscelli *
fscarscelli@geomar.de
GEOMAR Helmholtz Centre for Ocean Research Kiel
Kiel, Schleswig-Holstein, Germany

Claudius Zelenka*
cze@informatik.uni-kiel.de
Christian-Albrechts-University of Kiel
Kiel, Schleswig-Holstein, Germany

Daniyal Kazempour
dka@informatik.uni-kiel.de
Christian-Albrechts-University of Kiel
Kiel, Schleswig-Holstein, Germany

Peer Kröger
pkr@informatik.uni-kiel.de
Christian-Albrechts-University of Kiel
Kiel, Schleswig-Holstein, Germany

Florian Schütte
fschuette@geomar.de
GEOMAR Helmholtz Centre for Ocean Research Kiel
Kiel, Schleswig-Holstein, Germany

Figure 1: Deployment of a glider from the RV Meteor in the Tropical Altantic - Photo by Mario Müller (GEOMAR).

Abstract

Unsupervised anomaly detection (AD) on scientific sensor data is challenging due to the absence of labels, heterogeneous sequence lengths, and mixed anomaly types. We present a case study and an evaluation process for AD on multivariate data-sequences recorded by an autonomous ocean glider. The dataset consists of 341 variable-depth flights (3–920 m, 1 m vertical resolution) collected by glider IFM03 during R/V Meteor's cruises M105 and M106 in the Central Tropical Atlantic. We train a compact variational recurrent autoencoder with a CNN preprocessing layer and bidirectional LSTM encoder (16-dimensional latent) and a lightweight decoder. Reconstruction error (MSE) yield flight-level anomaly scores. In lieu of labels, we evaluate the AD system by means of oceanographic knowledge via expert inspection, separation of reconstruction error between normal flights and anomalies, and clustering of the learned embedding. Moreover, we present the result of the model in its geospatial context. Our contribution is a practical case-study of deploying unsupervised AD on real-world, unlabelled oceanographic data.

*Both authors contributed equally to this research.

GeoAnomalies '25, Minneapolis, MN, USA

ACM ISBN 979-8-4007-2260-8/2025/11
https://doi.org/10.1145/3764914.3770592

CCS Concepts

• **Computing methodologies → Anomaly detection**.

Keywords

Case Study, Real World Dataset, Autonomous Vehicles, Physical Oceanography, Microstructure Sensor, Anomaly Detection, Unsupervised Learning, Deep Learning, Data Visualization

ACM Reference Format:

Federico Scarscelli, Claudius Zelenka, Daniyal Kazempour, Peer Kröger, and Florian Schütte. 2025. Anomalies In The Depths: Evaluating Unsupervised Anomaly Detection on Real-World Unlabelled Data. A Case Study on Autonomous Ocean Gliders. In *The 2nd ACM SIGSPATIAL International Workshop on Geospatial Anomaly Detection (GeoAnomalies '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3764914.3770592

1 Introduction

Anomaly Detection (AD) is the task of finding anomalous or erroneous samples in data. This can be very useful in many sciences and applications, where measured data is necessary for empirical research and verification. In this paper we present a case study of the application and evaluation of unsupervised anomaly detection methods on oceanographic data. More specifically, we analyse the measurement of autonomous ocean gliders (more details below). These gliders are highly sensitive measuring platforms of ocean parameters and there is a large range of anomalies that can occur without categorization. The spurious nature of the anomalies makes this task entirely unsupervised, which means there are no labels and validation sets at all. Nevertheless, this data is used in downstream oceanographic models and research (at GEOMAR and other institutions), such as [8]. This makes the application and analysis of anomaly detection so important, but also so challenging.

Related work

The challenge of unsupervised anomaly detection evaluation. Ma et. al [11] emphasized in their work that unsupervised model selection in context of outlier detection is "notoriously difficult" due to the lack of validation data. Hence, the authors claim that the problem of unsupervised model selection for outlier detection is "vastly understudied" among the literature. Furthermore, they state that the state-of-the-art relies on internal model evaluation approaches, which nevertheless primarily depend on unlabeled input data and outlier scores that are individually based on strong assumptions.

Difference to internal evaluation of clusterings. Marques et. al [12] independently are in consensus with Ma et. al by stating that "the unsupervised evaluation of outlier detection results is still virtually untouched in the literature". The authors also oppose internal outlier detection validation to internal clustering validation by stating that while the former has been mostly overlooked in the literature, the latter has been applied and is mostly accepted as useful. Furthermore, the authors raise the claim that "It is important to acknowledge that no single index can capture all possible facets of the unsupervised outlier detection problem", a statement that underlines the relevance of this work in proposing an outlier detection evaluation measure capable of dealing with heterogenous sequence lengths and mixed anomaly types.

From evaluation of outliers in unsupervised settings to evaluation of outliers in time series and sequences. Getting more specific from the challenges of evaluating outlier detection results in unsupervised settings to anomaly detection in time series, Schmidl et. al. [16] investigated 71 outlier detection methods (supervised and unsupervised) on 976 time series datasets, accounting for differences in sequence lengths and type of anomaly detected. The authors found that most anomaly detection algorithms in the study exhibited high sensitivity with respect to their parameter settings. Notably, this work relies on external evaluation metrics and conducts the tests on single types of anomalies, not considering datasets exhibiting a mixture of different outlier types. Since neither of the discussed prior work elaborate on outlier detection evaluation on sequences exhibiting *heterogenous lengths*, *mixed types of anomalies*, in *unsupervised settings*, it is of utmost relevance to address these challenges within this work.

In the scientific literature there are plenty of works on anomaly detection [16]. Many that directly apply clustering [19] or apply statistical models [15] cannot be applied to our problem because glider flights have heterogeneous lengths and the anomalies are highly irregular. The improvement of Deep Learning models expanded our capabilities to compress the information. Beside making possible to reach performances in supervised tasks, learning expressive representations of complex data has proven a strong values also for AD purposes [14].

While statistical and shallow machine learning AD models try to identify anomalies via feature-engineering or analysing the statistical properties of the data, Deep learning anomaly detection methods typically leverage the reconstruction error or the latent space topology to find abnormal data [16]. This usually means that direct interpretation of the detection "rationale" is not possible, since the detection function is not based on the actual features of the data. Therefore, it is not easy to evaluate such methods when dealing with real-world, unlabelled datasets.

The focus of many recent works in the machine learning community has been on developing methods with strong performance on benchmark datasets [9]. Even though these methods are unsupervised they still use labels and metric like ROC for evaluation [1].

Hence in the absence of any labels, choosing the right method for a given real world task (that is not identical to existing benchmark settings) is very difficult, and requires an extensive evaluation of the results. A lot of downstream (in our case oceanographic) science depends on this data.

Contribution Therefore in our case study, we present an evaluation process and exemplary results that show how to work with completely unlabelled data. We discuss where the difficulties lie and, even though the task are complex and have no univocal solution, which tools and visualisation can actually help in a practical setting and where more research is necessary. In summary, the main contribution of our research is describing the evaluation of established unsupervised anomaly detection baselines in unlabelled real-world setting.

Autonomous Ocean Gliders While ship-deployed profilers (e.g. CTD rosettes) still represent the standard in seagoing physical observations, in recent years, autonomous ocean gliders have rapidly increased in popularity [13, 17]. They provide a flexible and convenient way to survey the waters beyond the course of the mothership, without requiring active attention from the researchers (beside deployment, course planning and retrieval), and being able to operate also with adverse sea and weather conditions [17]. Moreover, due to their size and good hydrodynamic properties, autonomous gliders have a minimal impact on the surrounding water compared to ship-based sensors, since the water displacement from the ship's hull and propellers interferers with the shallow water layers. Indeed, they represent the optimal platform to perform microstructure sensors surveys to collect uncontaminated small-scale turbulence data.

Within the limit of their battery autonomy, gliders can perform several "flights" during every deployment and could, theoretically, provide a continuos stream of data if organized into swarms (e.g. for environmental monitoring of a limited region). Processing and analysis of glider data is usually performed following "traditional" techniques that require the intervention of a skilled oceanographer. This disadvantage represents a strong limitation to the exploitation of available data, restricting survey design and data analysis to cruise/expedition level.

2 Data and Methods

Our dataset was produced by the "IFM03" glider, during its 10^{th} deployment, initiated during R/V Meteor's cruise M105 and retrieved during cruise M106 in the Central Tropical Atlantic [3–5]. The glider was equipped with 2 microstructure shear sensors mounted to the MicroRider instrument, in order to measure dissipation rate of turbulent kinetic energy. The dataset includes 341 variable-depth glider flights, ranging from 3 to 920 meters. The data has a resolution of 1 meter and include the variables listed in Table 1 together with geospatial metadata.

For our application, we consider every flight as a separate entity, without trimming them to a common depth value. The flights can be considered as a ordered collection of geospatially-referenced multidimensional sequence. Given $i = 1, 2 \dots n$ we can identify (u_i, v_i, t_i) as the geographical coordinates and timestamp of flight f_i, which consists in m data-points defined as follows:

$$f_i(h) = (x^1_{i,h}, x^2_{i,h}, \dots, x^p_{i,h})$$

where $h = 3, 4, \dots, m + 2$ is the depth in meters and x^k for $k = 1, 2 \dots p$ are p depth-varying variables (like those listed in Table 1). A graphical representation of a sample flight is illustrated in Figure 2. After interpolation, every flight is stored in a $(m \times p)$ tensor and standardized. Log-scaling is also applied to ϵ and N^2 before training, using the $x' = log(1 + x)$ transformation.

Table 1: Variables measured during the flights.

Variable Name	Measuring Unit	Label
Depth	m	dep
Dissipation rate (ϵ)	m^2/s^3	eps
Temperature	°C	t
Salinity		s
Stability (N^2)	s^{-2}	n2

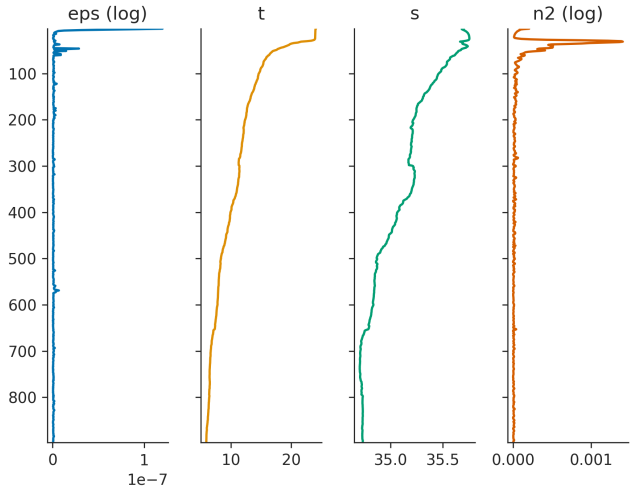

Figure 2: Depth profiles of data recorded during flight 210. The vertical axis measures the depth from the surface.

For our anomaly detection approach, we implement a LSTM-based Variational Recurrent Autoencoder (VRAE) with CNN enhancement [6, 7]. The model is designed to learn a compressed representation of the variable-length multivariate glider flights in a low-dimensional latent space and then reconstruct the original sequences from this representation. The variational approach should force the model to have a nicely structured latent space, where topological analysis the model embeddings is possible [2].

Because 341 sequences is actually very few data for deep learning, we need to strongly encourage the model to learn such a meaningful latent space. This intention is practically achieved in the model architecture also using a "smaller" decoder compared to the encoder. Therefore we use a small model, with most of the parameters in the encoder for good feature extraction, a very small latent space and a shallow decoder to prevent memorization. This forces strong encoder embedding performance rather than reconstruction fitting. Using a simpler decoder helps lowering the overall number of model parameters, while maintaining the same encoder complexity, also preventing overfitting.

The **encoder** consists of a convolutional preprocessing layer (3×3 kernel) and a bidirectional LSTM (128 hidden units). The final hidden states from both directions of the LSTM are concatenated and passed through two separate linear layers to produce the parameters of the latent distribution. This process maps each input profile to a 16-dimensional latent vector.

The **decoder**, made by a single-layer LSTM (48 hidden units), aims to reconstruct the original sequence from a latent vector z, sampled from the learned distribution. The output of the decoder LSTM at each step is passed through a final linear layer to reconstruct the original multivariate data point.

The model is then trained using the AdamW [10] optimizer (learning rate: $3{\times}10^{-4}$, weight decay: $5{\times}10^{-5}$) for 10 epochs with

a batch size of 32. We implement a beta-annealing schedule that gradually increases the weight of the KL-divergence term in the loss function to a maximum of 0.05, helping to avoid posterior collapse. Teacher forcing with a decay schedule is also employed to stabilize the decoder training.

To account for the different importance of variables, we apply weighted reconstruction loss with higher weights for turbulence measurements ($w_\epsilon = 2.0$).

After training, the VRAE should have learnt how to encode normal patterns of the glider flights into a compact latent representation and reconstruct them accordingly. Anomalies are expected to have a higher reconstruction error, as the model will be less effective at compressing and decompressing unseen or rare patterns. Indeed, the reconstruction loss can be used to identify possible anomalies: given a certain threshold λ, a flight f_i is labelled as anomaly if

$$||f_i - \hat{f}_i||_d > \lambda$$

where $||\cdot||_d$ is a suitable reconstruction error function. In absence of a validation set, the λ can be tuned based on statistical analysis of the reconstruction errors, using standard deviation rule or a target percentage of anomaly in the data. Also empirical strategies can be employed, like elbow-rule on the ordered reconstruction errors. Note that the anomaly detection error function $||\cdot||_d$ can differ form the training loss $||\cdot||_t$ using during model training. In our case, since the training loss function changes across epochs due to the beta-annealing schedule, we decided to employ MSE for anomaly detection. This is done in order to provide a coherent detection rule during the training, allowing for comparison and evaluation of the model's performances at different epochs.

The calculation of the MSE reconstruction error for flight f_i requires the computation of the $(j \times k)$ flight residuals matrix

$$E_{f_i} = (e_{jk})$$

as the matrix that contains the residuals for every predicted variable at every depth point $j = 1, 2, \ldots m$ of flight f_i as follows:

$$e_{jk} = (x_{ij}^k - \hat{x}_{ij}^k)^2$$

Given the along-depth sum of errors for variable k

$$\eta_k = \sum_{j=1}^{m} e_{jk}$$

we introduce the variable-specific anomaly score $\alpha_k(f_i)$ as the ratio between the sums of error in variable k and the total sum of errors:

$$\alpha_k(f_i) = \frac{\eta_k}{\sum E_{f_i}}$$

The value of $\eta_k(f_i) \in [0, 1]$ represents the "contribution" of variable k to the detection of f_i as anomaly and will be used to interpret the results of the AD system. Since distribution of the residuals e_{jk} can vary across variables (and can be different from a Gaussian), the $\eta_k(f_i)$ can be scaled dividing for its median across the i to help interpretation.

Since the primary focus of this work is evaluating AD systems applied to a real-world case study with unlabelled data, the results will be discussed following three different approaches: field-knowledge evaluation, statistical analysis of the system output and study of the embeddings produced by the deep learning model.

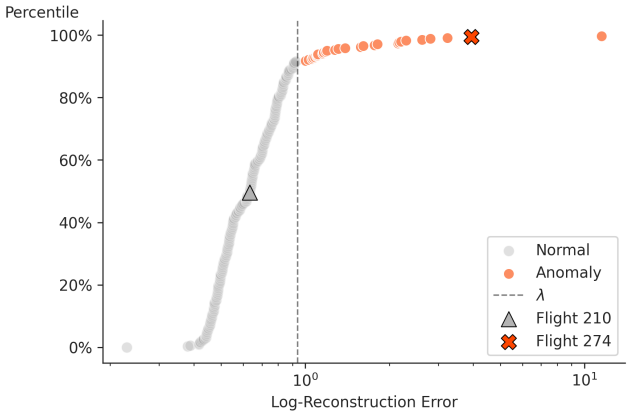

Figure 3: Ordered reconstruction errors obtained during the anomaly detection process. The errors of the provided examples of normal ab abnormal data are highlighted.

3 Results and Discussion

The inference reconstruction errors (Figure 3) are used to detect the abnormal sequences. We chose the value of λ analysing the ordered error curve in Figure 3, where a sharp knee is clearly visible, leading to identification of approximately 10% of anomalies.

When analysed with oceanographical knowledge, the flights recognised as positives (anomalies or novelties in the data) exhibit patterns related to rare physical conditions or measuring errors (Figures 5). The substantial difference within positive and negative group can be also appreciated by their statistical analysis in Figure 4.

The downstream "eyeball inspection" and the statistical characterization of the data is the bridge between our deep learning approach and the traditional ones (manual labelling and statistical AD). These passages are fundamental to initially assess the validity

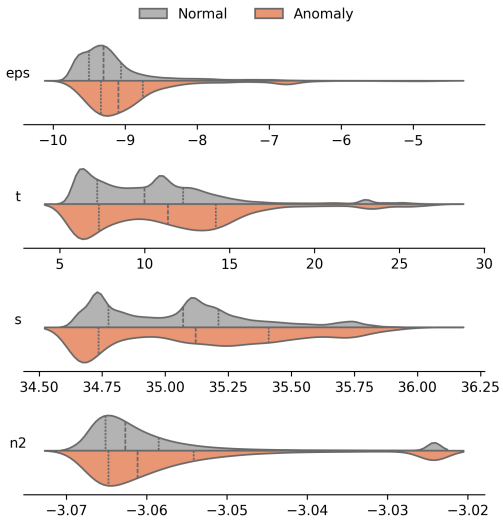

Figure 4: Statistical analysis of the detection results.

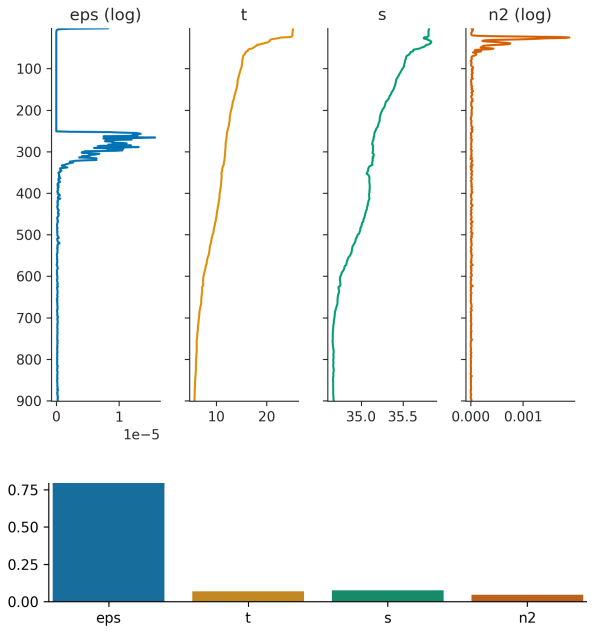

Figure 5: Depth profile of flight 274, an example of anomaly/novelty detected by the system (depth under the surface on vertical axis). The barplot in the bottom-figure represents the α anomaly scores for the flight.

of any AD model, but are even more important for autoencoder-based methods, due to the low interpretability of the inference process. No AD system can be considered trustworthy if it does not comply with this first evaluation stage, regardless of its technical integrity.

Employing expert validation *after* inference allows to dramatically reduce the effort required to the oceanographer in order to check the data quality. It is possible to argue that manual labelling of 341 sequence is possible in a reasonable time, yielding a possible validation or labelled training set. However, since this kind of measurements are affected by multiple geospatial factors (like seasonality, large and small scale systems, regional characteristics etc.), the representativeness of the labelled sample is difficult to assess, especially for anomalies and novelties. Due to this reason, our approach follows the idea to leave the model decide on the "easy" judgments (common data with low reconstruction error) and focus expert attention to unusual data. This is even more important if applied to real-time data processing during research cruises, where scientists' attention is a rare good that needs to be channelled into many different tasks.

Of course, this approach does not solve the problem of having undetected anomalies (false negatives) that could have a low reconstruction error while still featuring uncommon data. If this contingency is particularly relevant (e.g sensitivity of the system is crucial for security reason) it is always possible to lower the value of λ to meet the requirements.

To further increase the interpretability of the results, we use the variable-specific anomaly score $\alpha_k(f_i)$ to unveil the possible drivers of the detection of the flight as anomaly. In Figure 5 the anomaly scores for flight 274 (Figure 5) are presented. As we can see, the water profile exhibits a strong turbulent layer around 300 meters, quite unusual in this regional context. This reflects in the high value of α_ϵ, since the anomaly in the variable ϵ provokes more than the 80% of the total reconstruction error for this flight. This metric can help to understand what causes the detection of the anomalies, however, its practical use is not always easy. The ambiguity is caused by the difference in distribution of the variables' residuals, generating not consistent comparison between the anomaly scores, especially between variables with different order of magnitude like Dissipation rate/Stability and Temperature/Salinity.

To further evaluate the AD properties of our model, we analyse the latent space. Since a validation set is not available in our case, we tried to apply clustering to the embeddings, which can be seen as an alternative way to double-check the AD performance. Our idea is to observe how the detection-reconstruction error $||\cdot||_d$ relates to the topology of the learned representation of the data, in order to understand if the VRAE is able to capture the nature of our AD problem.

To visualize the latent space, we reduce its dimensions using t-SNE [18] and we use the reconstruction errors to color-map the embeddings (Figure 6). It is possible to see a clear left-right gradient in the representation of the latent space, however, since the visualizations obtained with t-SNE can be misleading sometimes [20], we try to investigate the real structure of the latent space via clustering on its full dimensions.

K-Means with an arbitrary k on the latent space gives us the results shown in Figure 7. Clearly this is not a perfect result, the highest reconstruction loss flights are not all together in one cluster,

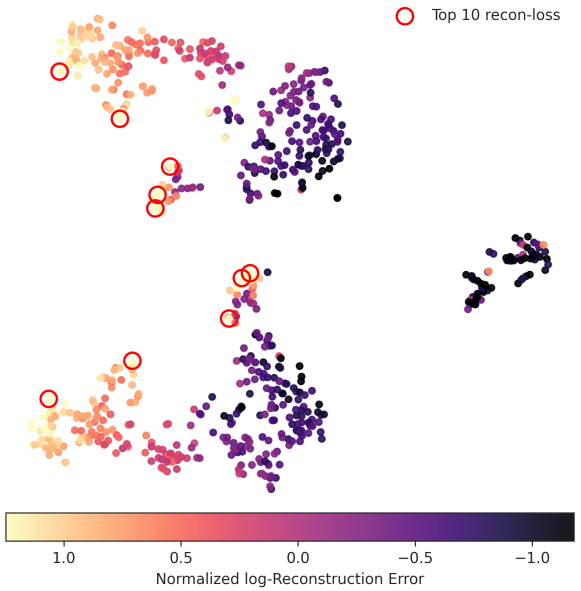

Figure 6: t-SNE representation of the latent space, coloured using a reconstruction error colour-map.

Figure 7: t-SNE representation of the latent space clustering.

but they are on the outside edge of clusters, and we can even see that the clusters are characterized by differences in the distribution of reconstruction error, as shown in Figure 8. This suggests that the insights from Figure 6 are correct and the probabilistic data-representation learnt by the VRAE reflects the concept of "normality" that we are interested to.

The detection of high-reconstruction-error sequences showed in Figure 3 leads to the identification of the geospatial anomalies in the dataset (Figure 9). Their distribution across the glider path does not seem random, featuring higher concentration of abnormal data around deployment (north-east) and retrieval (south-west) locations and also in another area along the course. While the anomalies around the extremes of the track are easily explainable by the ship's influence and/or the manoeuvrers (especially the ones directly before the retrieval), the profiles around 8°N can actually contain novelties or less common data. Unfortunately, understanding the cause of this anomaly "hotspot" is very complex, especially without going deep into the interpretation of the physical conditions of the area and perform an oceanographical study of the data. An ocean front, a storm or a seamount could have caused the mentioned anomalies and an oceanographer would easily understand the causal relationship if needed, however, we are interested in providing results agnostic to field-knowledge.

To further analyse the output of the system at 8°N, we focus on the α anomaly scores values along the path (Figure 10), in order to get variable-specific insights about what is inducing the anomaly detection. Of course, the α_k are strongly correlated along k and exhibit similar patters, however, given the high values of ϵ and N^2, apparently the water turbulence is driving the detection results. Naturally, these interpretation are based only on eyeball inspection of the plots and do not take into account any statistical testing or uncertainty estimate, nevertheless it can be interesting to spatially visualize the reconstruction residuals. Additional research can be

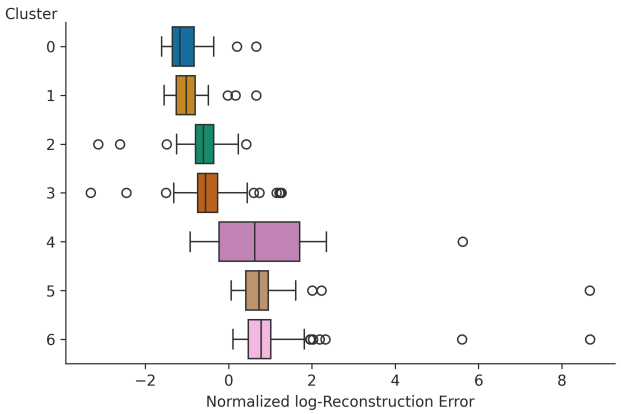

Figure 8: Boxplot of the reconstruction error across latent space clusters shown in Figure 7.

made in the direction of improving the spatial robustness of this approach, e.g. exploiting the points' neighbourhood information or applying more sophisticated methodologies such as Gaussian processes.

4 Conclusion

We presented a deep-learning anomaly detection task on 341 multivariate sequences from an autonomous ocean glider and proposed a practical evaluation process for real-world, unlabelled datasets.

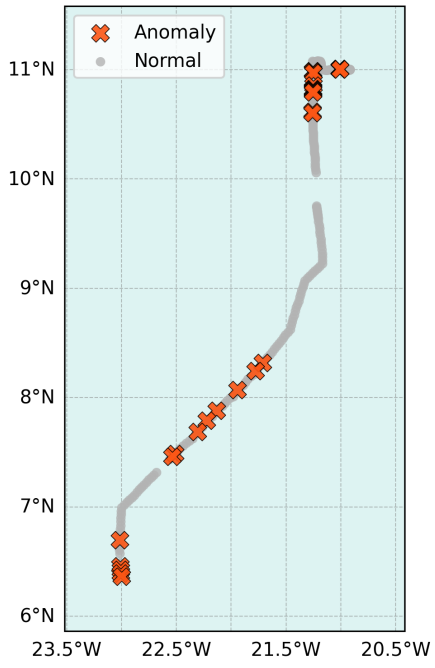

Figure 9: Geospatial visualization of the anomalies along the glider path.

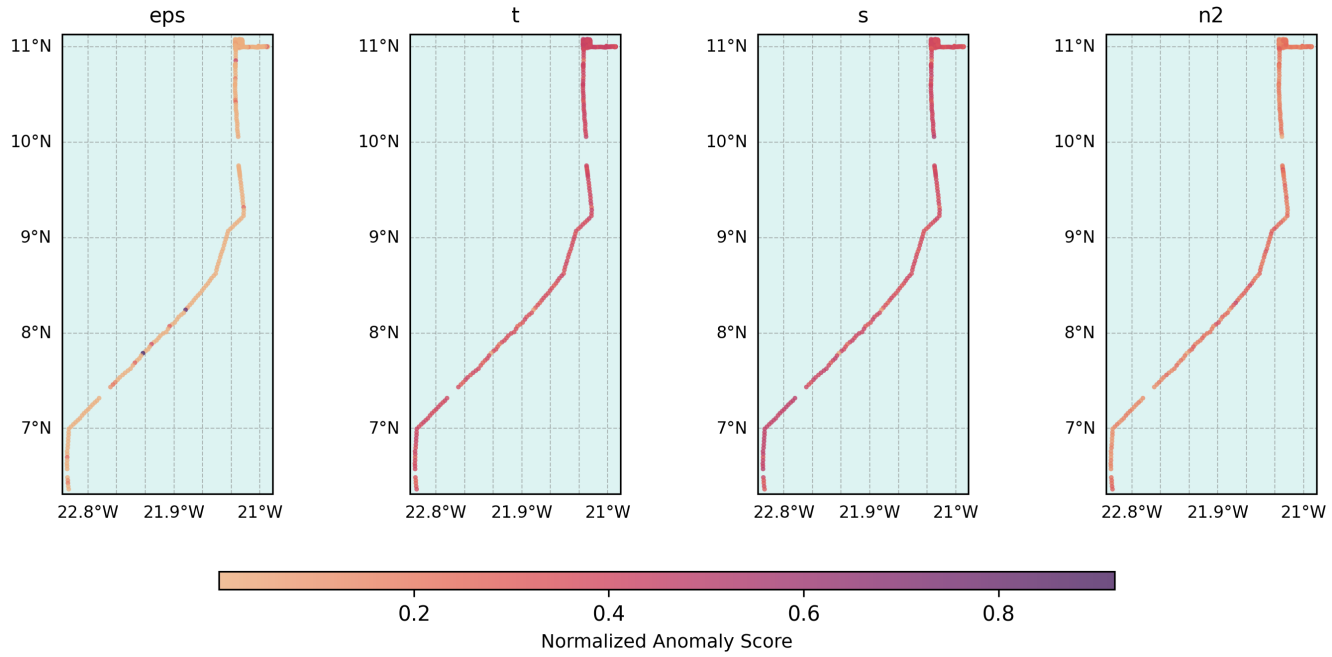

Figure 10: Geospatial visualization of the normalized α anomaly scores.

The challenges in this case study include small-sample training, lack of ground truth, threshold sensitivity and variable length data.

We implemented a compact VRAE with CNN preprocessing layer and bidirectional LSTM encoder that yields reconstruction-based flight and variable-level anomaly scores. We discussed the evaluation of the anomaly detection system through field-knowledge checks, statistical description of normal and anomalous sets and topological analysis of the learned embeddings.

The results jointly indicate the latent space organizes profiles by "normality", providing some evidence about internal validity, indicating that deep-learning anomaly detection can be successful even in such challenging conditions. Of course, many aspects of the problem still need an exhaustive answer: how to tune the system without human intervention or semi-supervised workflows, how to improve the latent representation of the anomalies and how to interpret the result in relation to their geospatial context.

Moreover, further efforts are required to evaluate the system by increasing the size and variety of the dataset, including measurements from different regions and seasons. This would also enable a discussion on the external validity of the results. Making the system available to a growing number of oceanography researchers and incorporating their feedback will, in turn, generate new insights through its dissemination.

Acknowledgments

The first Author is supported by Helmholtz School for Marine Data Science (MarDATA), Helmholtz Association.
This work was supported by Helmholtz AI computing resources (HAICORE) of the Helmholtz Association's Initiative and Networking Fund through Helmholtz AI.
This work was supported by the GEOMAR Innovation Call 2025.
The photo in the first page was taken by Mario Müller (GEOMAR).

References

[1] Maxime Alvarez, Jean-Charles Verdier, D'Jeff K. Nkashama, Marc Frappier, Pierre Martin Tardif, and Froduald Kabanza. 2022. A Revealing Large-Scale Evaluation of Unsupervised Anomaly Detection Algorithms. *ArXiv* abs/2204.09825 (2022).

[2] Sungtae An, Shenda Hong, and Jimeng Sun. 2020. Viva: semi-supervised visualization via variational autoencoders. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 22–31.

[3] Marcus Dengler. 2020. Microstructure data from a MicroRider/Glider package (METEOR cruise M105 and M106). doi:10.1594/PANGAEA.920597 Artwork Size: 145020 data points Pages: 145020 data points.

[4] GEOMAR. 2014. Meteor M105. https://www.geomar.de/en/research/expeditions/detail-view/exp/324217?cHash=0aad09059a363b6fffd1a55dba082e3d

[5] GEOMAR. 2014. Meteor M106. https://www.geomar.de/en/research/expeditions/detail-view/exp/325623?cHash=fa85b666869eff9ca2b7324e743f9297

[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735

[7] Diederik P. Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat] doi:10.48550/arXiv.1312.6114

[8] Nicolas Kolodziejczyk, Pierre Testor, Alban Lazar, Vincent Echevin, Gerd Krahmann, Alexis Chaigneau, Claire Gourcuff, Malick Wade, Saliou Faye, Philippe Estrade, Xavier Capet, Laurent Mortier, Patrice Brehmer, Florian Schütte, and Johannes Karstensen. 2018. Subsurface Fine-Scale Patterns in an Anticyclonic Eddy Off Cap-Vert Peninsula Observed From Glider Measurements. *Journal of Geophysical Research: Oceans* 123, 9 (2018), 6312–6329. doi:10.1029/2018JC014135

[9] Andreas Lohrer, Darpan Malik, Claudius Zelenka, and Peer Kröger. 2024. GADformer: A Transparent Transformer Model for Group Anomaly Detection on Trajectories. arXiv:2303.09841 [cs] doi:10.48550/arXiv.2303.09841

[10] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Bkg6RiCqY7

[11] Martin Q Ma, Yue Zhao, Xiaorong Zhang, and Leman Akoglu. 2023. The need for unsupervised outlier model selection: A review and evaluation of internal evaluation strategies. *ACM SIGKDD Explorations Newsletter* 25, 1 (2023), 19–35.

[12] Henrique O Marques, Ricardo JGB Campello, Jörg Sander, and Arthur Zimek. 2020. Internal evaluation of unsupervised outlier detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14, 4 (2020), 1–42.

[13] Ali Mashayek, Nick Reynard, Fangming Zhai, Kaushik Srinivasan, Adam Jelley, Alberto Naveira Garabato, and Colm-cille P. Caulfield. 2022. Deep Ocean Learning of Small Scale Turbulence. *Geophysical Research Letters* 49, 15 (Aug. 2022), e2022GL098039. doi:10.1029/2022GL098039

[14] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. 2022. Deep Learning for Anomaly Detection: A Review. *Comput. Surveys* 54, 2 (March 2022), 1–38. doi:10.1145/3439950

[15] Peter J Rousseeuw and Mia Hubert. 2018. Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 2 (2018), e1236.

[16] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment* 15, 9 (2022), 1779–1797.

[17] Louis St. Laurent and Sophia Merrifield. 2017. Measurements of Near-Surface Turbulence and Mixing from Autonomous Ocean Gliders. *Oceanography* 30, 2 (2017), 116–125. https://www.jstor.org/stable/26201858 Publisher: Oceanography Society.

[18] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using T-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.

[19] Zhuo Wang, Yanghui Zhou, and Gangmin Li. 2020. Anomaly detection by using streaming K-means and batch K-means. In *2020 5th IEEE international conference on big data analytics (ICBDA)*. IEEE, 11–17.

[20] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. 2016. How to Use t-SNE Effectively. *Distill* (2016). doi:10.23915/distill.00002

Received 15 August 2025; accepted 26 September 2025