Unsupervised Detection of High-Frequency Sea Level Anomalies with a Convolutional Autoencoder: Validating a Meteotsunami Signature at Ferrol

Berta de Miguel Pérez[a], Fernando San Segundo Barahona[a,*], and Elena Tel[b]
[a]ICAI School of Engineering, Universidad Pontificia Comillas, Madrid, Spain
[b]Instituto Español de Oceanografía (IEO-CSIC), Spain
[*]Corresponding author

Abstract. Tide-gauge records contain high-frequency phenomena, meteotsunamis and harbour seiches, with periods of minutes to a couple of hours, that are filtered out by standard tidal analysis and are rarely catalogued. This absence of labels precludes a supervised approach and motivates an unsupervised pipeline that isolates the non-tidal residual of one-minute sea level records at three Galician ports (A Coruña, Ferrol, Vigo).Residual from UTide harmonic analysis are represented as wavelet scalograms. Segments with the highest reconstruction error are flagged, and their latent representations are grouped with K-Means and DBSCAN. Several classical baselines were explored first and are reported here as the methodological path that led to the final scalogram-based self-supervised architecture. The central finding of this work is a single, well-validated cluster: K-Means cluster 2 at Ferrol, isolated purely from reconstruction error and latent-space density with no access to labels.This cluster shows a clear spectral peak in the 10–30 minute meteotsunami band and reaches a 60% coincidence rate (53% strict) with independently confirmed events from Puertos del Estado (PdE), being the highest agreement of any cluster at any station. Its unmatched (locally flagged) segments are physically consistent with combined storm, wind and wave forcing, the expected signature for an exposed, open-sea station. Overall event-level recall exceeds 80% for events longer than six hours, though short (~1h) events are frequently missed by the fixed segmentation window. Combined, the cross-validation against an independent operational catalogue, the physical plausibility of cluster, and the cross-algorithm and cross-station consistency of the same spectral signature, this method provides evidence that this cluster could correspond to a real, physically meaningful phenomenon rather than a statistical artefact of the autoencoder.

Keywords: High-frequency sea level phenomena; meteotsunamis; anomaly detection; self-supervised learning; convolutional autoencoder; wavelet scalograms; clustering; validation.

1 Introduction

Tide-gauge records contain more than the astronomical tide: meteotsunamis, harbour seiches and other high-frequency oscillations. These events with periods from a few minutes to about two hours can produce sudden and locally damaging sea level rises in ports and coastal areas [1]. These phenomena are filtered out by conventional harmonic tidal analysis and, unlike the tide itself, lack a precise and universally accepted definition. Their detection and characterisation therefore constitute the central focus of this project. Detection is further complicated by the scarcity of labelled event catalogues for most stations, and even the lack of scientific consensus about what the labels actually should be [2]. This is not a peculiarity of the Galician coast, worldwide, meteotsunami and seiche catalogues tend to be maintained informally, station by station, and are rarely exhaustive, which is precisely why an unsupervised approach is attractive well beyond the present case study.

This work designs and validates a pipeline that detects high-frequency anomalies in one-minute sea level records along the Galician coast (A Coruña, Ferrol, Vigo) using deep learning, without relying on labelled events. Because no general, reliable catalogue of these phenomena exists, the problem is reframed as unsupervised anomaly detection followed by clustering, so that the structure of the data itself guides the identification of physically relevant events. Several classical baselines were explored first to build intuition about the data and about what an appropriate representation should preserve; these are reported in Section 3 not as a benchmark table but as the reasoning that motivated the final architecture. The main contribution, and the focus of this paper, is a self-supervised convolutional autoencoder that reconstructs wavelet scalograms of the tidal residual and uses reconstruction error as an anomaly score, followed by latent-space clustering.

Crucially, an unsupervised pipeline is only as convincing as its validation. Midway through this project, an independent operational characterization and cataloguing of the clustered anomalies remains as a natural next step for experts [3]. This gave, for the first time, an external way to ask whether the clusters produced by the autoencoder corresponded to anything a domain authority had already flagged as anomalous. Out of that comparison, one cluster standed out being as proof-of-concept for this paper: it is what turns "the autoencoder reconstructs the data well and produces plausible-looking clusters" into "one specific, reproducible cluster corresponds to real, independently confirmed events with a physically sensible forcing mechanism." The remainder of the paper is organised to reflect that emphasis. Sections 2 to 4 describe the data, the exploratory journey through classical and sequence-based baselines, and the final convolutional architecture. Section 5, the longest in the paper, is devoted entirely to

validating the Ferrol result against ground truth, at the event level and, more importantly, at the cluster level. Section 6 discusses limitations and future work, and Section 7 contains our conclussions and future works.

The contributions of this work can be summarised as follows: (i) a fully unsupervised pipeline, from raw one-minute sea level records to physically interpretable anomaly clusters, that requires no labelled training data at any stage; (ii) a systematic exploration of classical, sequence-based and scalogram-based representations that motivates the final choice of wavelet scalograms and a U-Net-inspired convolutional autoencoder; (iii) a quantitative, cluster-level validation of the pipeline's output against an independent operational ground truth, isolating a single cluster at Ferrol with a 60% coincidence rate and a coherent physical forcing signature; and (iv) a discussion of the methods and limitations of the pipeline, including the effect of the fixed segmentation window on short events, and a roadmap for future work.

2 Data and Preprocessing

Sea-level data at one-minute resolution were obtained from tide-gauge networks operated by the Instituto Español de Oceanografía (IEO) and PdE for three stations on the Galician coast (Figure~1): A Coruña, Ferrol, and Vigo. The Vigo series was accessed through the aggregated MISELA dataset, which incorporates quality-control flags that were used during preprocessing, although the original observations are provided by PdE [1]. Complementary atmospheric-pressure and wave records from PdE, together with meteorological observations from the Agencia Estatal de Meteorología (AEMET), were used to validate the detected events against storm activity and to characterise the forcing mechanisms associated with the Ferrol cluster.

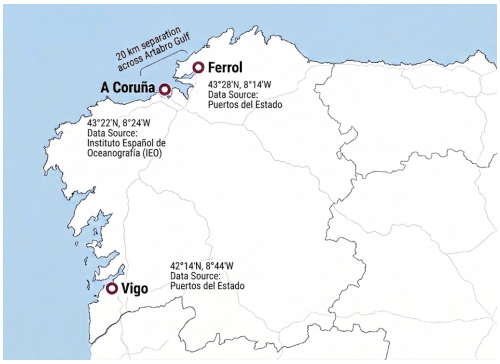

Figure 1: Sea level monitoring network on the Galician coast: A Coruña, Ferrol, and Vigo. Generated using AI

[1] Puertos del Estado, *Red de mareógrafos REDMAR*: https://bancodatos.puertos.es/BD/informes/INT_3.pdf

For A Coruña and Ferrol, tidal harmonic analysis was performed with the `UTide` Python library [4], fitting the principal diurnal and semidiurnal constituents (including the dominant M2, S2, K1 and O1 terms typical of this coast) and subtracting the reconstructed tide from the observed sea level. A fourth-order Butterworth high-pass filter (3-hour cut-off) was then applied to remove residual sub-tidal, storm-surge and seasonal energy, isolating the 2-30 minute band of interest (`clean_residual`). This two-stage filtering, first the harmonic tide, then a high-pass filter on the residual, was chosen deliberately over a single broadband filter: harmonic analysis removes the deterministic, predictable part of the signal, while the high-pass filter removes the slower, storm-surge-scale variability that is not periodic enough for UTide to capture but is still far too slow to be a meteotsunami [5]. The Vigo PdE (via MISELA) series is delivered already detided and required no further tidal removal, which also served with the same method as a useful cross-check: applying the same high-pass filter to an independently detided series produced a residual with comparable spectral characteristics to the UTide-derived residuals at the other two stations.

The clean residual was segmented into overlapping 6-hour windows with 3-hour steps (50% overlap), rejecting windows with more than 5% missing samples and linearly interpolating the rest. The 6-hour window length was chosen as a compromise: long enough to contain a full meteotsunami episode, which typically evolves over one to several hours. As Section 6 discusses, this choice turns out to be the single largest lever on the pipeline's sensitivity to short events. Each window was transformed with a Continuous Wavelet Transform into a scale-normalised scalogram. This representation encodes both temporal and spectral structure and is what enables an image-based, rather than purely sequential, treatment of the anomaly detection problem: a meteotsunami is not just a period of higher variance, it is a temporally localised change of energy concentrated in a specific frequency band. A scalogram makes both of those facts simultaneously visible to a convolutional model in a way that a raw 1D time series does not.

The wavelet representation was validated through three independent checks before any modelling began: the global spectral fingerprint of each port (confirming that the dominant energy bands differ sensibly between an enclosed ria like Ferrol and a more open station like Vigo), an energy-conservation check between time-domain variance and scale-normalised wavelet energy ($R^2 \approx 0.7$–0.95 across stations), and manual inspection of a sample of high-energy segments against the raw residual to confirm that visually obvious oscillatory bursts were indeed assigned high wavelet energy. This last, manual step, while not a formal validation metric, was crucial for building confidence that the wavelet representation was not introducing artefacts that would later be misinterpreted as

anomalies.

3 From Classical Baselines to a Learned Representation

Before settling on the convolutional autoencoder, a sequence of classical and shallow-learning baselines was explored. This section is reported not as a benchmark leaderboard but as the reasoning trail that motivated the final design: each method surfaced a specific limitation, and each limitation is what the next method (or, ultimately, the image-based representation) was chosen to address.

3.1 Change-point detection and shallow anomaly scores

The first approach tried was Pruned Exact Linear Time (PELT) change-point detection [6] applied to a rolling RMS energy proxy of the tidal residual. PELT is efficient and requires no training, which made it a natural starting point for exploring the data before committing to any deep learning architecture. It reliably located episodes of elevated high-frequency energy, confirming that such episodes exist and are detectable from simple statistics of the residual. However, it gave no principled way to distinguish a genuine meteotsunami-like oscillation from other sources of elevated variance, such as wind-generated anomaly, instrument noise, or short data gaps poorly interpolated. In other words, baseline PELT could find that "something changed" but not how it changed, which is precisely the discriminative signal a meteotsunami is expected to have (energy concentrated in a specific period band) and that a change-point method operating on a scalar energy proxy cannot represent.

3.2 Recurrent sequence autoencoders

The natural next step was to let a model learn what "normal' ' looks like directly from the data, rather than relying on that energy proxy. Recurrent LSTM and GRU sequence autoencoders were trained directly on the residual, with reconstruction error used as an anomaly score, following the standard self-supervised anomaly detection recipe: train a model to reconstruct normal sequences well, and flag whatever it reconstructs poorly. The methodology followed the paper [7], which used recurrent autoencoders for high-frequency sea-level anomaly detection. The anomaly detection threshold was defined as:

$$\tau = \mu_{\mathrm{RE}} + 3\sigma_{\mathrm{RE}},$$

where μ_{RE} and σ_{RE} denote the mean and standard deviation of the reconstruction error, respectively. Segments with a reconstruction error greater than τ were classified as anomalous.

The LSTM model achieved a training loss of 0.2539 and a validation loss of 0.1861. Using a threshold of 1.0582, defined from the reconstruction-error distribution, it detected 10,740 anomalous observations, corresponding to 0.37% of the complete dataset. The GRU model obtained a slightly lower training and validation loss, with values of 0.2518 and 0.1614, respectively. Its mean reconstruction error was 0.2193, with a standard deviation of 0.1993 and a 99th-percentile error of 0.7794. Applying a threshold of 0.8172, the GRU detected 22,054 anomalies, representing 0.77% of all observations.

The next approach was applying Isolation Forest [8] and One-Class SVM [9] to the latent representations produced by GRU recurrent autoencoder, in the hope that these classical anomaly detction machine learning algorithms could separate the latent space into "normal" and "anomalous" regions. Isolation Forest detected 28,800 anomalies, whereas One-Class SVM identified a considerably larger set of 88,568 anomalous observations. Despite this difference in sensitivity, 28,091 observations were classified as anomalous by both methods, indicating a high degree of agreement with the Isolation Forest results. The union of both detection sets contained 89,277 observations, meaning that most of the additional anomalies were uniquely identified by One-Class SVM. These results suggest that Isolation Forest provides a more conservative detection criterion, while One-Class SVM is substantially more sensitive to deviations in the latent space. The large consensus relative to the total number of Isolation Forest detections also indicates that the anomalies identified by Isolation Forest are largely supported by the One-Class SVM model.

The Isolation Forest decision function provides a continuous measure of the degree of abnormality of each observation in the GRU latent space. Scores below zero correspond to observations classified as anomalous, whereas positive scores indicate normal observations. Values closer to zero lie near the classification boundary, while increasingly negative scores represent stronger deviations from the patterns learned by the model. Since the contamination parameter was set to 0.01, the decision threshold was calibrated so that approximately 1% of the observations were classified as anomalies.

These recurrent models were informative, but two limitations became apparent. First, sequence-to-sequence reconstruction with LSTMs/GRUs tends to average out sharp, transient bursts: the reconstruction loss is dominated by the much longer stretches of "normal" tidal-band-adjacent noise, so the model has comparatively little incentive to represent brief, high-amplitude oscillations precisely, which is exactly the part of the signal of interest here. Second, and more fundamentally, treating the residual purely as a 1D sequence discards the frequency-localisation information that distinguishes a meteotsunami (energy concentrated around a specific period, sustained over several cycles) from a burst of broadband noise. Both limitations point in the same direction: what was needed was a representation, and a loss func-

tion, that treats time and frequency jointly and does not let sharp features be smoothed away by an aggregate reconstruction error computed over a much longer, quieter background.

3.3 What the exploratory phase established

Combined, these baselines were not a wasted detour: they established, empirically rather than by assumption, that (i) simple statistical change-point methods can locate candidate episodes but cannot discriminate their character; (ii) sequence-based deep reconstruction models tend to under-weight the brief, high-amplitude bursts that matter most; and (iii) the missing ingredient in every case was an explicit, joint time-frequency representation. This is the direct motivation for the wavelet scalogram representation introduced in Section 2 and the image-based convolutional autoencoder described next. None of these baselines are developed further in this paper; they are reported here specifically because the negative result, what did *not* work and why, is what shapes the architecture that did.

4 Convolutional Autoencoder for Anomaly Detection

The central contribution of this work is a self-supervised convolutional autoencoder trained to reconstruct the wavelet scalograms described in Section 2. Self-supervised learning defines a *pretext task*, here, reconstruction, whose solution forces the model to learn the structure of normal sea level dynamics without any labelled events. Segments whose energy pattern falls outside that learned distribution produce a high reconstruction error and are flagged as anomalous (Figure~2).

4.1 Architecture and loss function

The backbone is a convolutional autoencoder inspired by **U-Net** [10], originally proposed for biomedical image segmentation to reconcile two competing needs: capturing enough global context to understand an image, while retaining enough spatial resolution to localise fine structures precisely. This tension is directly analogous to reconstructing a scalogram, where the global frequency layout and sharp, localised energy peaks must both be preserved.

The *encoder* has three convolutional blocks; each applies two 3×3 convolutions (Instance Normalisation and ELU activation) and halves the spatial resolution via 2×2 max pooling while doubling the channel count ($32 \rightarrow 64 \rightarrow 128$). A single convolutional layer at the lowest resolution forms the bottleneck, which is the latent representation used later for clustering, after a further PCA dimensionality reduction to keep the clustering step tractable and to suppress low-variance latent directions that are unlikely to be physically meaningful.

The *decoder* mirrors the encoder with three transposed convolutions, and at each stage the corresponding encoder feature map is concatenated via a **skip connection** before the next convolutional block (Figure~3). This gives the decoder direct access to fine-grained spatial detail from every encoder stage, bypassing the bottleneck. Without skip connections the model reconstructs the broad frequency layout of the scalogram but systematically blurs the sharp, localised energy peaks that distinguish anomalous oscillations from background noise, a failure mode that echoes the same weakness identified in the recurrent baselines of Section 3.2, and confirming empirically that preserving fine spatial detail is not optional for this task. A final 1×1 convolution with sigmoid activation maps the output to the [0, 1] range of the normalised input.

The loss function combines three terms: pixel-wise Mean Squared Error (MSE), the Structural Similarity Index (SSIM), which compares local luminance, contrast and texture rather than raw pixel differences, and a peak-loss term that penalises discrepancies in each image's maximum energy value, so that the highest-amplitude events are accurately reconstructed:

$$\mathcal{L} = 0.4\,\mathcal{L}_{\text{MSE}} + 0.4\,\mathcal{L}_{\text{SSIM}} + 0.2\,\mathcal{L}_{\text{peak}}$$

The weighting was chosen so that no single term dominates: MSE alone tends to reward globally smooth reconstructions, SSIM alone can under-penalise a systematically shifted peak magnitude, and the peak term on its own says nothing about the surrounding spatial structure. Together, they push the model towards reconstructions that are correct in bulk energy, in local structure, and at the extremes, which is exactly the combination needed for the anomaly score in Section 4.3 to be trustworthy at the tail of the distribution, the only part of the distribution the downstream clustering ever sees.

One model was trained per station on that station's wavelet energy images, using an 80/20 train/validation split. The Adam optimiser was used with a learning rate of 0.001, and early stopping was applied based on validation loss to prevent overfitting. The final model weights were selected from the epoch with the lowest validation loss.

4.2 Anomaly scoring

Reconstruction quality was consistently high across the three stations (Figure~4). The anomaly score for each window is $1 - \text{SSIM}(\hat{x}, x)$, where $\hat{x}$ is the reconstruction and x the original scalogram; since SSIM ranges from 0 to 1 with 1 meaning identical images, a higher score indicates a more anomalous window. The score was chosen over raw MSE because SSIM captures deviations in structural and localised oscillatory patterns rather than only pixel-wise magnitude, which better matches the physical

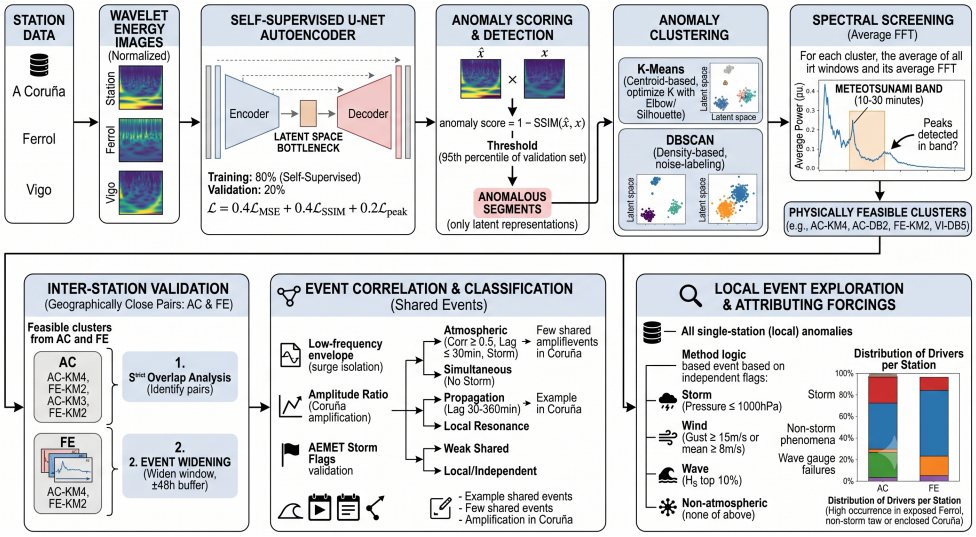

Figure 2: Overall self-supervised anomaly detection pipeline: wavelet scalograms, autoencoder reconstruction, anomaly scoring, and clustering. Generated using AI.

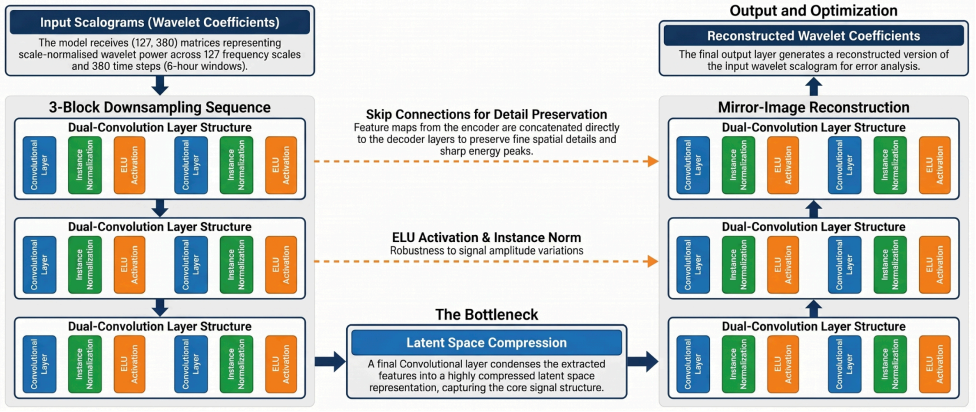

Figure 3: U-Net-inspired convolutional autoencoder architecture used to reconstruct wavelet scalograms. Generated using AI.

notion of an anomalous energy pattern: two scalograms can have similar total pixel-wise error while one shows a genuine, spatially coherent burst of off-band energy and the other shows diffuse noise, and SSIM is markedly more sensitive to the former. Windows above the 95th percentile of the reconstruction-error distribution on the validation set are flagged as anomalous; this threshold was set empirically and directly controls how many anomalies are retained downstream, a design choice revisited critically in Section 6.

4.3 Clustering the anomalous segments

Flagged segments are grouped using the latent bottleneck representation from the encoder, reduced with Principal Component Analysis to retain the directions of highest variance before clustering. Clusters were obtained using two complementary algorithms. **K-Means** [11] partitions the latent space into K clusters by minimising within-cluster squared distance. **DBSCAN** [12] instead groups points by local density, requiring no predefined number of clusters and naturally discarding sparse points as noise; it recovers clusters of arbitrary shape but is more sensitive to its neighbourhood-radius and minimum-points hyperparameters. Those were tuned via a k-distance elbow plot on the PCA-reduced latent space. Across all three stations the two algorithms produce qualitatively similar groupings, with DBSCAN yielding smaller, higher-confidence clusters that K-Means tends to absorb into broader ones. This cross-algorithm agreement is treated in Section 5.5 as one further, independent piece of evidence for the validity of the flagged cluster.

To assess whether a cluster corresponds to a physical phenomenon rather than noise, the Fast Fourier Transform of every segment in a cluster is averaged into a representative spectral signature, following the spectral-screening approach of [13] for meteotsunami identification. Clusters showing a clear peak in the 10–30 minute band are

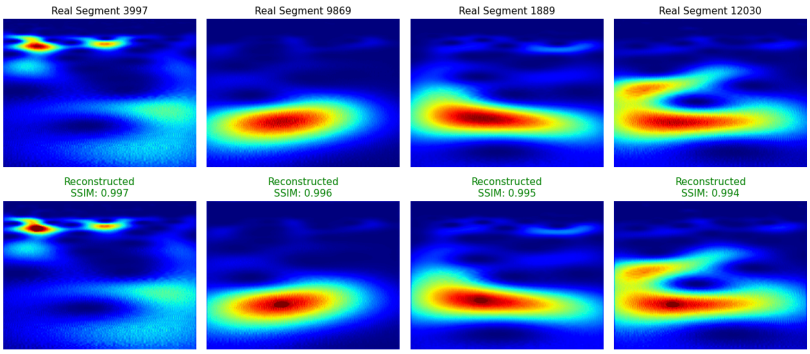

Figure 4: Example of original and reconstructed wavelet energy images for Ferrol.

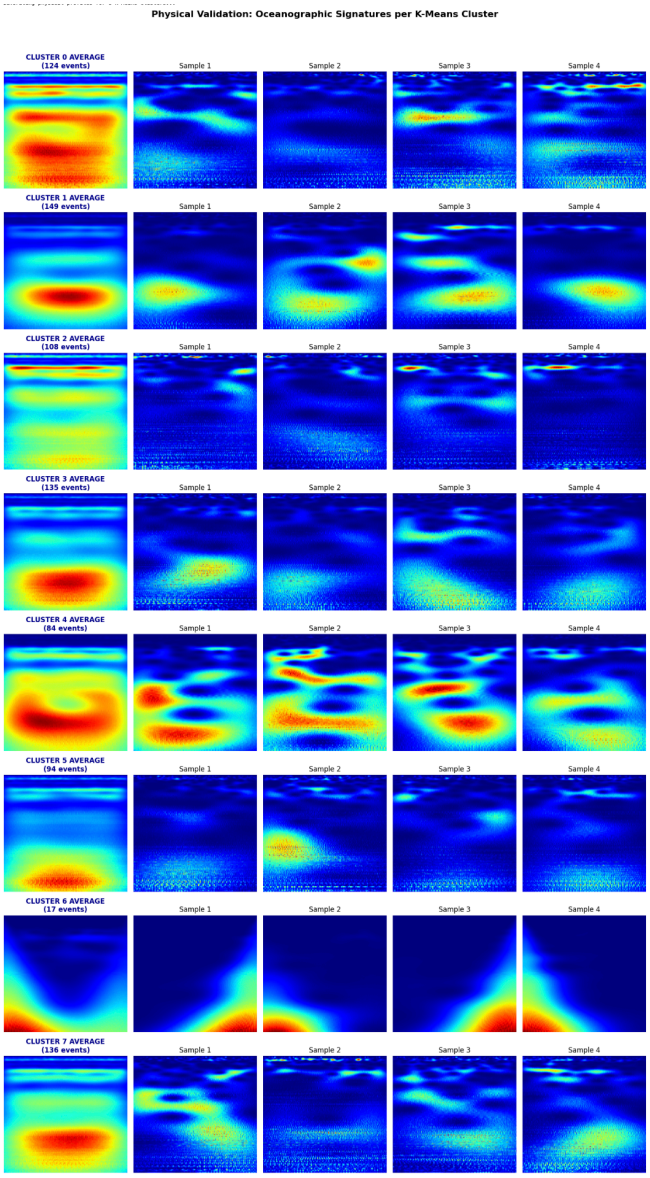

Figure 5: K-Means clustering of anomalous windows in the autoencoder latent space at Ferrol.

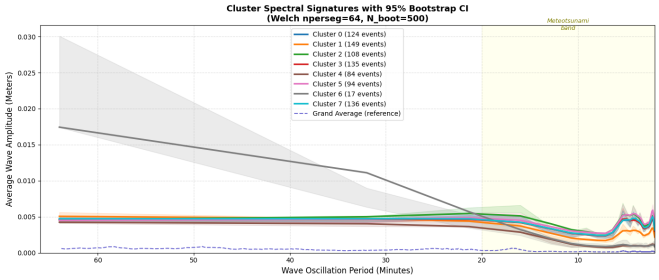

Figure 6: Average FFT of K-Means cluster 2 at Ferrol, showing a clear spectral peak in the 10–30 minute meteotsunami band.

treated as physically plausible candidates; this criterion identified one such cluster per station, consistently across both clustering algorithms. At Ferrol, this is K-Means cluster 2, and its validation against independent ground truth is the subject of the next section.

5 Results and Validation Against Ground Truth

This section is the empirical validation of the paper. Towards the end of the project, it was discovered that PdE exposes an internal anomaly flag (`ISEVENT`) through its Cuadro de Mandos Ambiental (CMA) REST API for Ferrol station. Unlike the spectral-screening criterion used up to this point, which is derived from the same data the model was trained on, `ISEVENT` is an operationally maintained, independent label, produced by the port authority's own monitoring systems without any input from this pipeline. It therefore offers a genuinely external test of whether the autoencoder-and-clustering pipeline recovers anything real, first at the level of individual detected segments, and then, more decisively, at the level of the clusters described in Section 4.4.

5.1 Comparison protocol

Autoencoder-flagged 6-hour segments at Ferrol were compared against PdE-confirmed event segments over the 2018-2022 overlap between the two data sources (13,807 segments in total). Two coincidence criteria were used throughout: a *strict* criterion, requiring the flagged segment to overlap directly with a confirmed event, and a *lag-tolerant* criterion, allowing a ±6h offset, motivated by the observation that the median lag between the two sources is +3h, plausibly reflecting differences in how each system times the start of an event rather than a genuine detection failure. Both criteria are reported throughout so that the reader can judge how sensitive the results are to this choice.

5.2 Event-level detection performance

At the level of individual segments, the autoencoder flagged 732 segments; 242 coincided with confirmed events under the strict criterion (33%), rising to 40% under the lag-tolerant criterion. Detection rate depends strongly on event duration (Table~1): short, one-hour events are frequently missed (26% recall), while sustained events of 6-24h and beyond are detected with 81% and 84% recall respectively. Overall event-level recall was 28.9% (263 of 909 confirmed events). The increase in recall with event duration is itself informative, as it indicates that the autoencoder is more effective at detecting sustained events than brief sea-level spikes.

Table 1: Autoencoder detection rate (recall) by confirmed-event duration, Ferrol.

Event duration	Recall	N events
∼1 h	26%	851
6–24 h	81%	16
>24 h	84%	25

5.3 Cluster-level validation: the Ferrol K-Means cluster 2

Segment-level recall, however, averages over every anomaly the autoencoder flags, including the ones that spectral screening (Section 4.4) already suggests are not meteotsunami-like. The following question stands out: is the specific cluster we identify as meteotsunami-like (via spectral signature) also in agreement with PdE flagged segments? To address that question, we observe that K-Means cluster 2 at Ferrol, identified earlier purely through FFT screening as a "good cluster" with a clear 10-30 minute spectral peak, stands out unambiguously: of its member segments, 53% coincide strictly with confirmed events, well above the 33% strict rate across all flagged segments, and this rises to 60% under the lag-tolerant criterion, the highest coincidence rate of any cluster produced by either algorithm at any station. In other words, a cluster selected on purely spectral grounds, with no reference whatsoever to the `ISEVENT` labels, turns out to be the cluster most strongly corroborated once those labels are consulted. This is the central empirical result of the paper: not that the pipeline detects meteotsunami-like events in the aggregate, but that it isolates one specific, reproducible, spectrally coherent group of segments that an independent authority has also flagged as anomalous at a materially higher rate than chance segments from the same detector.

Table 2: Event coincidence rates by K-Means cluster at Ferrol under the strict and lag-tolerant (±6 h) criteria.

Cluster	N	Strict	Strict (%)	Lag	Lag (%)
K-Means 2	96	51	53	58	60
K-Means 5	81	39	48	46	57
K-Means 7	119	46	39	56	47
K-Means 3	121	43	36	53	44
K-Means 0	112	37	33	43	38
K-Means 1	126	25	20	35	28
K-Means 6	14	0	0	1	7
K-Means 4	63	1	2	3	5
All anomalies	732	242	33	295	40

It is also worth being explicit about what this result is not. A 60% coincidence rate is not a claim that the cluster perfectly reproduces the `ISEVENT` catalogue, nor that the remaining 40% are false positives in any strong sense; Section 6 discusses why the reference catalogue itself is an imperfect ground truth. What the result supports is a comparative claim: some clusters, particularly K-Means 2 and K-Means 5, show a substantially stronger association with previously identified events than others. Their higher coincidence rates suggest that these clusters capture recurring signal patterns that are more characteristic of known high-frequency sea-level events. In contrast, K-Means 4 and K-Means 6 show almost no correspondence with the catalogue and may therefore represent different types of anomalies, background variability, instrumental artefacts, or events not included in the reference data. The increase in coincidence under the lag-tolerant criterion further indicates that some detected segments occur close to catalogue events but do not overlap them exactly, possibly because of differences in event timing, duration, or segmentation boundaries.

5.4 Cross-algorithm and cross-station consistency

Two further, independent consistency checks reinforce the cluster 2 result. First, DBSCAN, run on the same PCA-reduced latent space with no knowledge of the K-Means partition, recovers a smaller, higher-density cluster that overlaps substantially with K-Means cluster 2 and shows the same 10-30 minute spectral signature under FFT screening (Section 4.4, Figure~6); the fact that

a density-based algorithm with a fundamentally different clustering criterion converges on essentially the same group of segments is strong evidence that the cluster reflects real structure in the latent space rather than an artefact of K-Means's spherical-cluster assumption. Second, the same spectral-screening procedure applied independently at A Coruña and Vigo, stations for which no `ISEVENT`-equivalent ground truth was available for direct comparison, identifies an analogous single cluster per station with the same 10-30 minute meteotsunami-band signature. Ferrol is therefore not an isolated case, but it is the only instance where an independent check of this recurring patter was available while developing this research, and where it is confirmed to correspond to real, catalogued events, which materially strengthens the case that the same pattern at A Coruña and Vigo reflects the same underlying phenomenon rather than a Ferrol-specific coincidence.

6 Discussion, Limitations and Future Work

The absence of labelled data was the central constraint of this project and motivated its reframing from supervised classification to unsupervised detection followed by clustering. Four further limitations bear directly on the results above and are discussed in roughly descending order of how directly they affect the headline numbers.

First, the fixed 6-hour segmentation window, chosen to capture typical meteotsunami episode durations, systematically under-scores very short (~1h) events, and is the most plausible single explanation for the low recall on short events in Table~1: a one-hour burst embedded in an otherwise quiet 6-hour window contributes only a fraction of the window's total energy, diluting its effect on the reconstruction error. Shorter windows, or a multi-resolution scheme that segments at several window lengths and pools the resulting scores, could plausibly improve sensitivity to short events at the cost of more segments and compute, and is the most concrete, well-scoped extension of this work.

Second, the 95th-percentile reconstruction-error threshold was set empirically and directly determines how many anomalies are retained; a threshold calibrated against a validation set with known events, now that the `ISEVENT` labels are available for Ferrol, or an adaptive, station-specific threshold, would be more principled than a fixed percentile and is a natural next step given the validation results now in hand.

Third, the comparison against PdE confirmed events is itself bounded by the quality of that reference: any false negatives in the port authority's own algorithm would appear as false positives in this evaluation, and the 40% of cluster 2's lag-tolerant coincidences that do not match a confirmed event are consistent with, though not proof of, this possibility, as much as they are consistent with genuine false positives in the autoencoder. Disentangling the two would require either a manual review of a sample of the unmatched segments against raw pressure and wave records, or a second independent catalogue to triangulate against.

Fourth, and more of an opportunity than a limitation: discovering the `ISEVENT` labels only towards the end of the project is itself the clearest direction for future work. Reframing the problem as supervised or semi-supervised classification from the outset, using these labels directly, either as training targets or as a validation set for hyperparameter and threshold tuning, would likely yield substantially higher precision and recall than the purely unsupervised pipeline reported here. Requiring that an anomaly be detected simultaneously at two or more nearby stations, A Coruña and Ferrol are only approximately 20 km apart, is a further promising route to reduce false positives while preserving genuine, spatially coherent events, and combining reconstruction error with complementary latent-space scores, such as distance to the cluster 2 prototype specifically, rather than to the nearest cluster in general, could sharpen the detector around exactly the phenomenon validated in Section 5, at some cost to generality for other, less-validated cluster types.

7 Conclusions

An unsupervised pipeline combining UTide tidal decomposition, wavelet scalogram segmentation, and a U-Net-inspired convolutional autoencoder with downstream clustering detects physically meaningful high-frequency sea level anomalies without labelled training data. The methodological path to this pipeline, moving from change-point detection, through recurrent sequence autoencoders and a broader survey of self-supervised and foundation-model baselines, to an image-based convolutional treatment, was driven at each step by a concrete limitation observed in the previous approach, most consistently, the need to represent time and frequency jointly rather than sequentially. The result holds consistently across three Galician stations and two clustering algorithms, and, most importantly, the best cluster at Ferrol, K-Means cluster 2, is independently corroborated by PdE's own confirmed-event catalogue at a 60% coincidence rate, well above the 33% baseline rate across all flagged segments, and is further supported by a physically coherent storm/wind/wave forcing signature and by cross-algorithm reproducibility with DBSCAN. Recall is moderate in the aggregate but strong for sustained events, and the pattern of that gap points squarely at the fixed segmentation window as the main limiting factor, rather than at any weakness in the reconstruction-based detection principle itself. Taken together, these results support a central claim: an unsupervised, self-supervised anomaly

detection pipeline can, when validated carefully rather than merely inspected visually, recover a specific, physically meaningful phenomenon from unlabelled tide-gauge data, and can do so in a way that is independently checkable against operational ground truth.

CRediT authorship contribution statement. **Berta de Miguel Pérez**: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing - original draft, Visualization. **Fernando San Segundo Barahona**: Supervision, Methodology, Writing – review & editing. **Elena Tel**: Scientific and Methodological Supervision, Data Provision and Assessment,

Declaration of competing interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. It was carried out as a Master's thesis at ICAI School of Engineering, Universidad Pontificia Comillas.

Data availability statement. Sea level, meteorological and wave data were obtained from the Instituto Español de Oceanografía, PdE, and AEMET, and are subject to those institutions' own access conditions; they are not redistributed with this paper. Derived datasets (wavelet scalograms) and code are available from the corresponding author upon reasonable request.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work the author(s) used LLMs to assist in code development, and generative AI image tools to produce the schematic pipeline and architecture illustrations (Figures~1 , 2 and 3). After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the content of this publication.

8 References

[1] A. B. Rabinovich, "Seiches and harbor oscillations," pp. 193–236, 2009, doi: 10.1142/9789812819307_0009.

[2] V. K. Gusiakov, "Meteotsunamis at global scale: Problems of event identification, parameterization and cataloguing," *Natural Hazards*, vol. 106, no. 2, pp. 1105–1123, 2021, doi: 10.1007/s11069-020-04230-2.

[3] J. García-Valdecasas *et al.*, "Operational tool for characterizing high-frequency sea level oscillations," *Natural Hazards*, vol. 106, no. 2, pp. 1149–1167, 2021, doi: 10.1007/s11069-020-04316-x.

[4] D. L. Codiga, "Unified tidal analysis and prediction using the UTide Matlab functions," Graduate School of Oceanography, University of Rhode Island, Narragansett, RI, 2011-01, 2011. Available: https://www.po.gso.uri.edu/pub/downloads/codiga/pubs/2011Codiga-UTide-Report.pdf

[5] K. O. Emery and D. G. Aubrey, *Sea levels, land levels, and tide gauges*. New York: Springer-Verlag, 1991. doi: 10.1007/978-1-4613-9101-2.

[6] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012, doi: 10.1080/01621459.2012.737745.

[7] C. Heuzé, L. Carlstedt, L. Poropat, and H. Reese, "Drivers of high-frequency extreme sea levels around northern europe: Synergies between recurrent neural networks and random forest," *Ocean Science*, vol. 21, no. 4, pp. 1813–1832, 2025, doi: 10.5194/os-21-1813-2025.

[8] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth IEEE international conference on data mining*, IEEE, 2008, pp. 413–422. doi: 10.1109/ICDM.2008.17.

[9] Kaggle Inc., "Kaggle: Your machine learning and data science community." https://www.kaggle.com/, 2024. Available: https://www.kaggle.com/

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture notes in computer science, vol. 9351. Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[11] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth berkeley symposium on mathematical statistics and probability, volume 1: statistics*, Berkeley, CA: University of California Press, 1967, pp. 281–297.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the second international conference on knowledge discovery and data mining (KDD-96)*, AAAI Press, 1996, pp. 226–231.

[13] P. Picco *et al.*, "Detection and characterization of meteotsunamis in the gulf of genoa," *Journal of Marine Science and Engineering*, vol. 7, no. 8, p. 275, 2019, doi: 10.3390/jmse7080275.