Advanced Data-Driven methods for marine environment dynamics prediction

Giovanni Ragusa

Dipartimento di Ingegneria

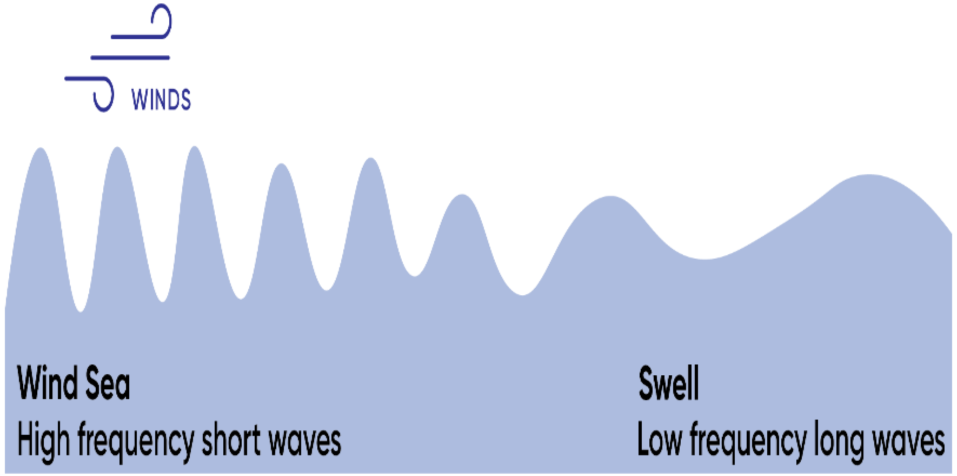
Università degli Studi di Messina

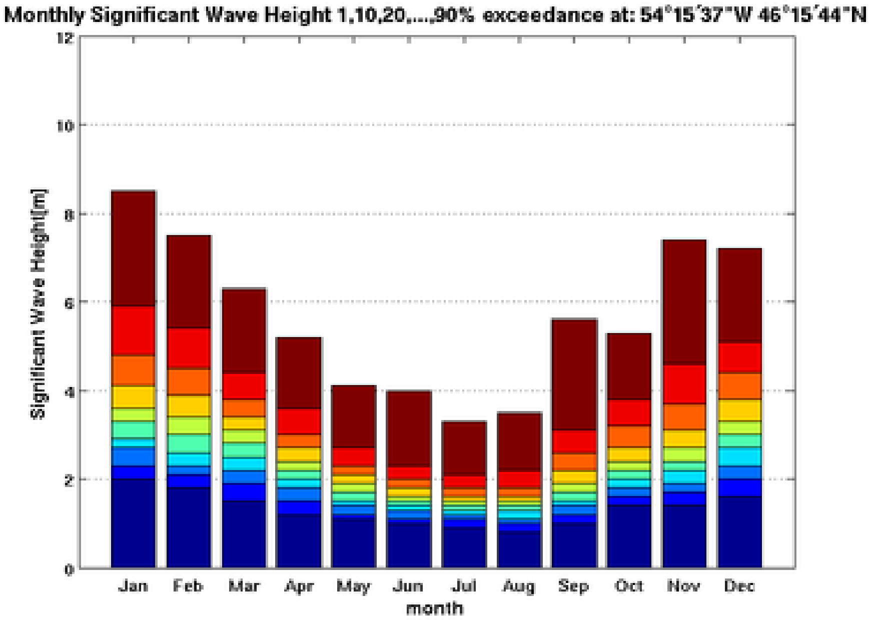
Tutor: Prof.ssa Maria Gabriella Xibilia, Prof.ssa Carla Lucia Faraci

Significant Wave Heights (HS) Prediction

Significant Wave Height (Hs) prediction is critical for:

- Maritime safety and navigation
- Coastal infrastructure design
- Offshore operations planning




Monthly Significant Wave Height 1,10,20,...,90% exceedance at: 54°15′37"W 46°15′44"N

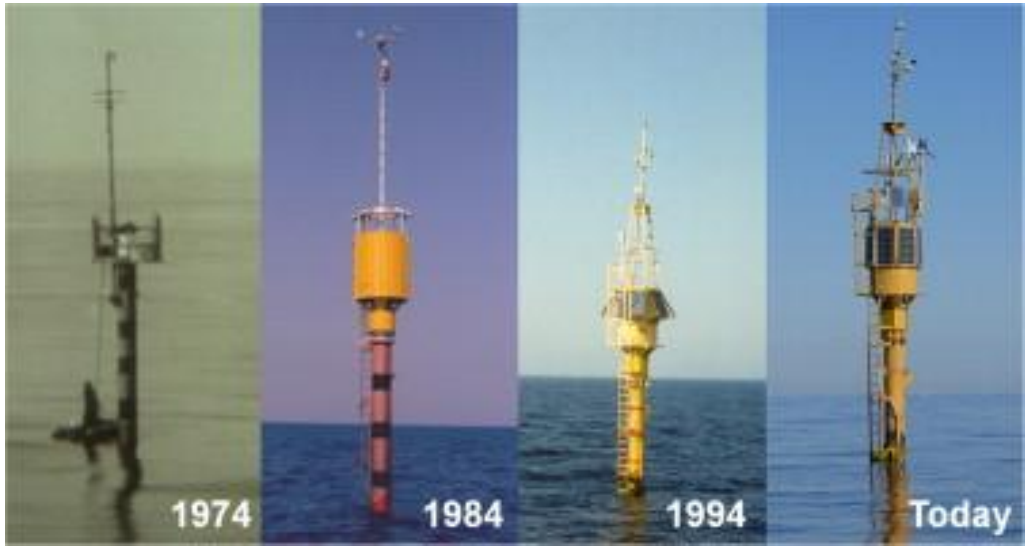
Motivation

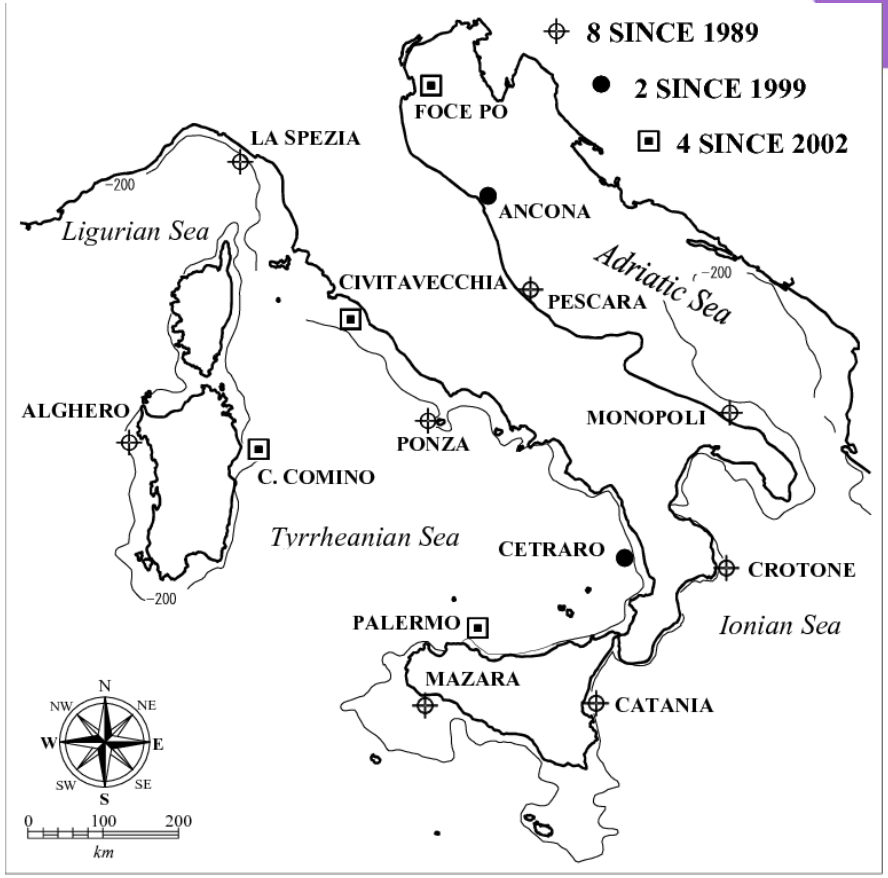
- The availability of **wave climate data** is important information for designing **offshore and coastal works**
- The main source of data comes from measurements made by buoys that are part of **sea monitoring networks** managed by national and international centres
- **The acquired data are often not continuous in time** due to damages and/or maintenance of measuring instruments
- The **quality** of data needs to be improved

Oil rig in storm is a painting by Ceri Jones

Italian Sea Monitoring Network

- The Italian Sea Monitoring Network, including **15 buoys** located in the deep water around the coast, shows an interruption period from 2014 to 2021. After 2021 only seven locations were restarted



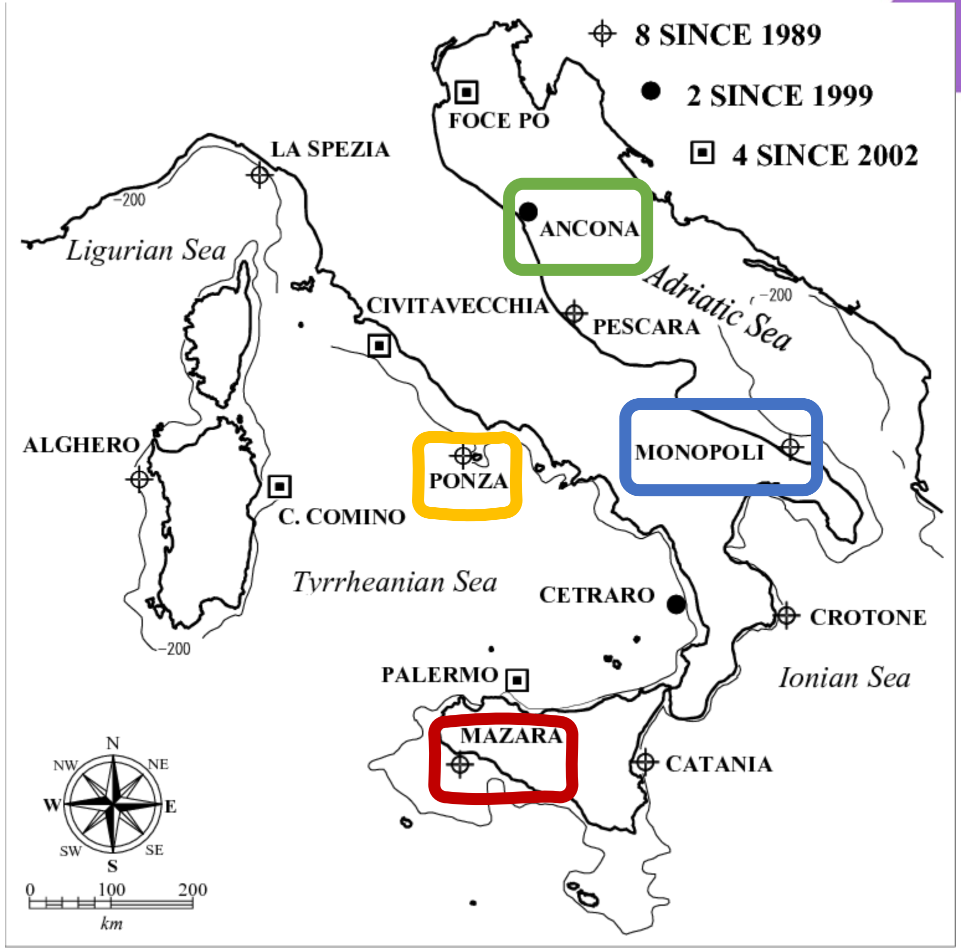


[1] Elisa Canepa, Sara Pensieri, Roberto Bozzano, Marco Faimali, Pierluigi Traverso, Luigi Cavaleri, The ODAS Italia 1 buoy: More than forty years of activity in the Ligurian Sea, Progress in Oceanography, Volume 135, 2015, Pages 48-63, ISSN 0079-6611

Dataset

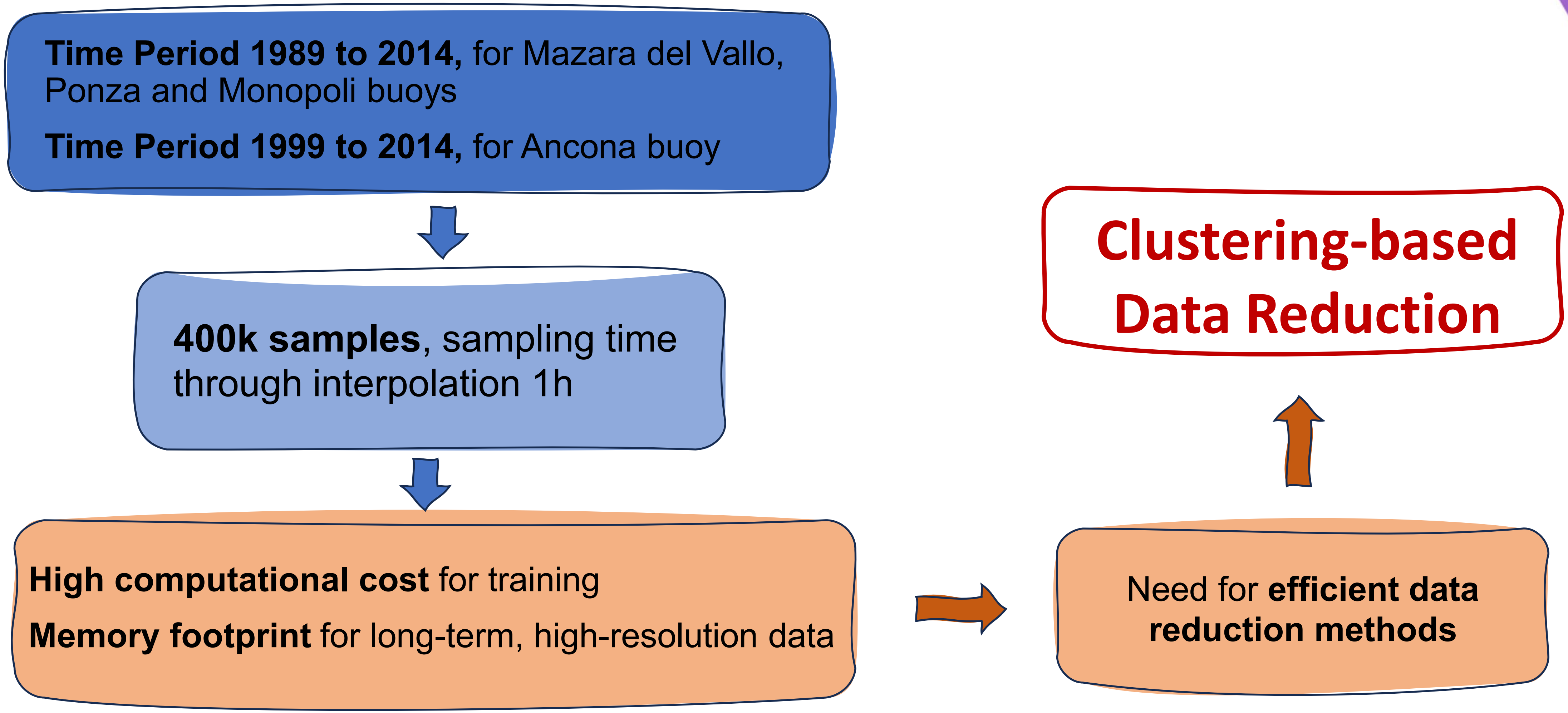
- **The buoys** of Mazara del Vallo and Ponza belong to the **Italian National Wave Recording Network RON** managed by the Agency for Environmental Protection and Technical Services, ISPRA.

  - ✓ **Mazara del Vallo** is located 13 km from the coast at a depth of 100 m.
  - ✓ **Ponza** is located 1.3 km from the coast at a depth of 115 m.
  - ✓ **Monopoli** is located 6 km from the coast at a depth of 85 m.
  - ✓ **Ancona** is located 30 km from the coast at a depth of 70m



[1] Elisa Canepa, Sara Pensieri, Roberto Bozzano, Marco Faimali, Pierluigi Traverso, Luigi Cavaleri, The ODAS Italia 1 buoy: More than forty years of activity in the Ligurian Sea, Progress in Oceanography, Volume 135, 2015, Pages 48-63, ISSN 0079-6611

Dataset



Time Period 1989 to 2014, for Mazara del Vallo, Ponza and Monopoli buoys

Time Period 1999 to 2014, for Ancona buoy

400k samples, sampling time through interpolation 1h

High computational cost for training

Memory footprint for long-term, high-resolution data

Need for **efficient data reduction methods**

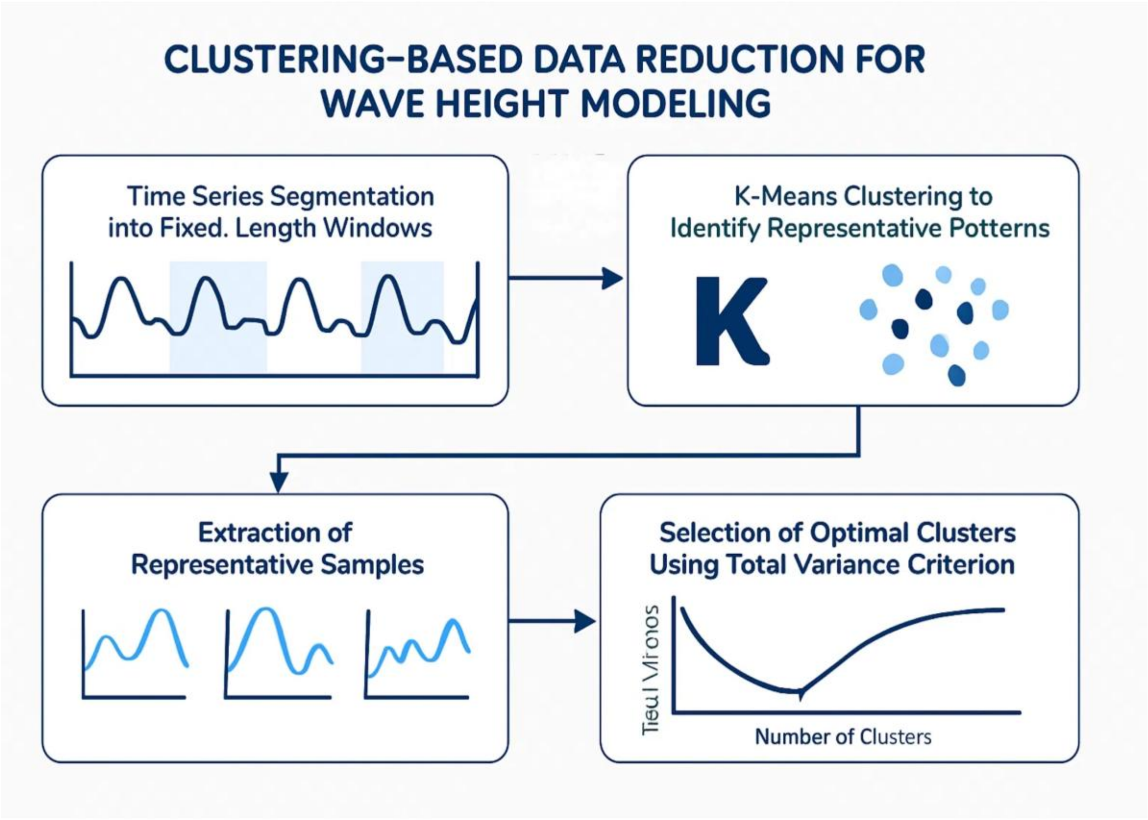
Clustering-based Data Reduction

Proposed Methodology

SECTION 2

Clustering-based data reduction framework

1. **Time series segmentation** into fixed-duration windows (B = 72 hours)
2. Application of **K-means algorithm** to identify representative patterns
3. Selection of optimal number of clusters via **Total Variance criterion** and **elbow method**
4. Extraction of a **representative subset** covering at least 15% of the original temporal span with approximately uniform distribution.



CLUSTERING-BASED DATA REDUCTION FOR WAVE HEIGHT MODELING

Time Series Segmentation into Fixed. Length Windows

K-Means Clustering to Identify Representative Potterns

Extraction of Representative Samples

Selection of Optimal Clusters Using Total Variance Criterion

Data Reduction with Clustering

K-means Algorithm:

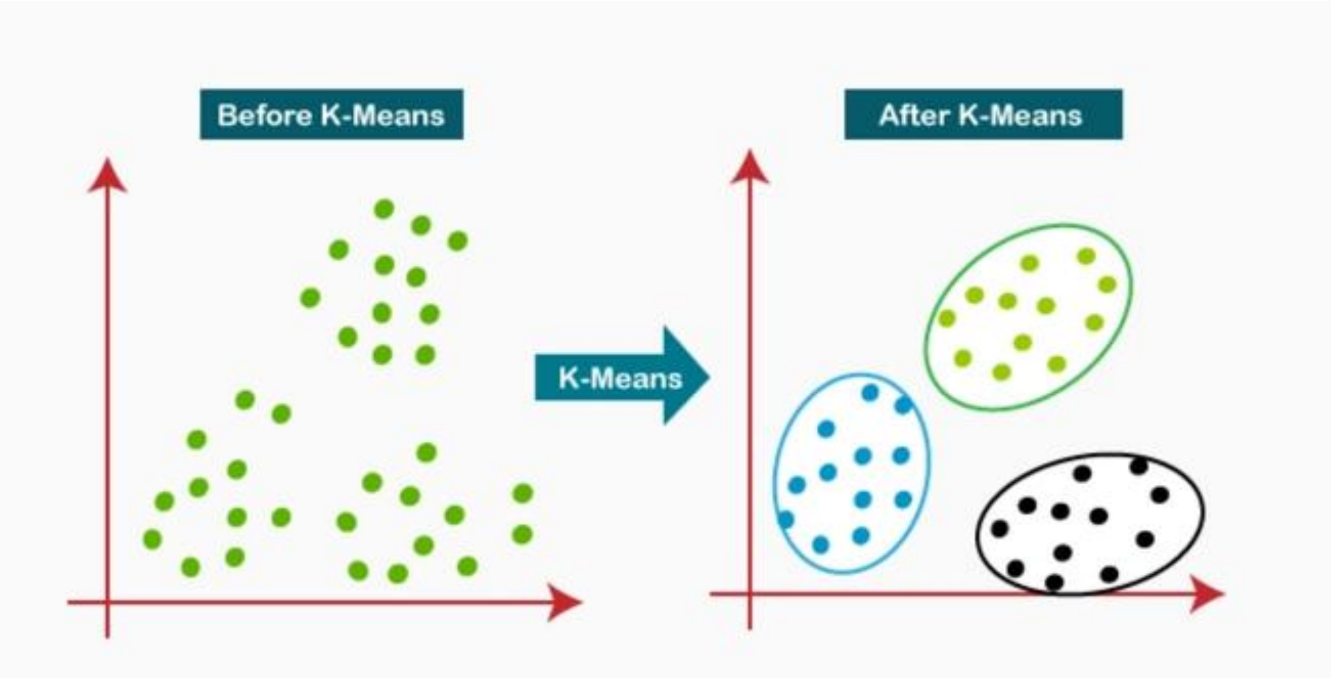
- Partitions data into k representative clusters
- Uses Square Euclidean Distance as metric:

$$D_{SE}(x,y) = \Sigma_{i=1}^{n} (x_i - y_i)^2$$

Optimal Number of Cluster Selection:

- Total Variance (TV) Criterion
- Elbow method with ERR TV = 0.05 threshold

$$TV(k) = \Sigma_{i=1}^{k} \Sigma_{x_j \in C_i} ||x_j - \mu_i||^2$$



Advantages of the approach

- Significant **reduction in computational load**
- Preservation of essential **wave climate dynamics**
- Capture of key events such as **storms and calm conditions**

Mazara del Vallo Example

Representative Subset Extraction:

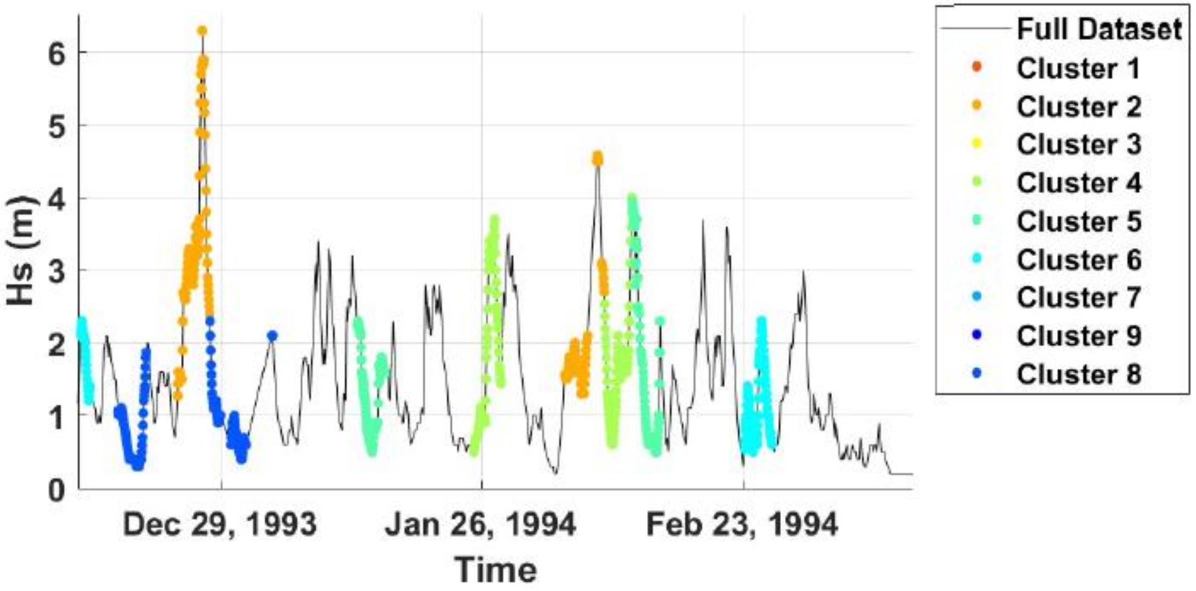
- Covers at least 15% of the original temporal span

Dataset analysis:

- Optimal number of clusters identified: $k_{opt} = 9$

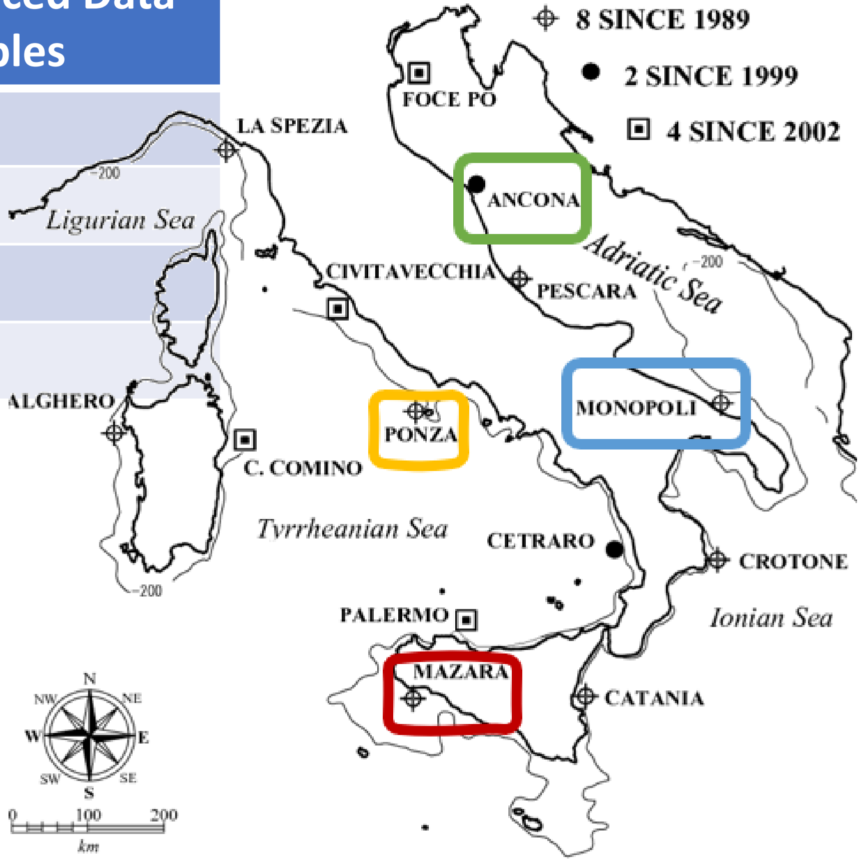
Cluster Characteristics:

- Clusters 1-3: High intensity storm events (Hs > 3m)
- Clusters 4-6: Moderate sea conditions
- Clusters 7-9: Calm sea conditions

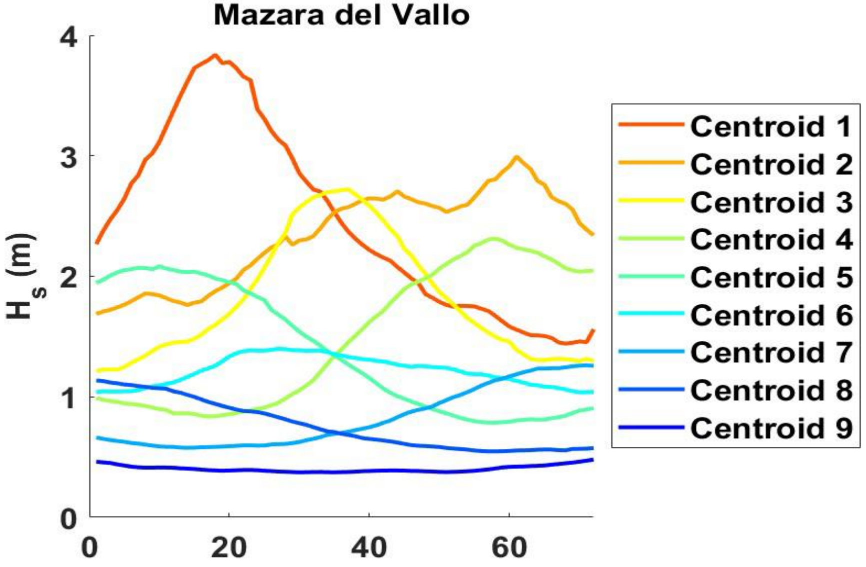


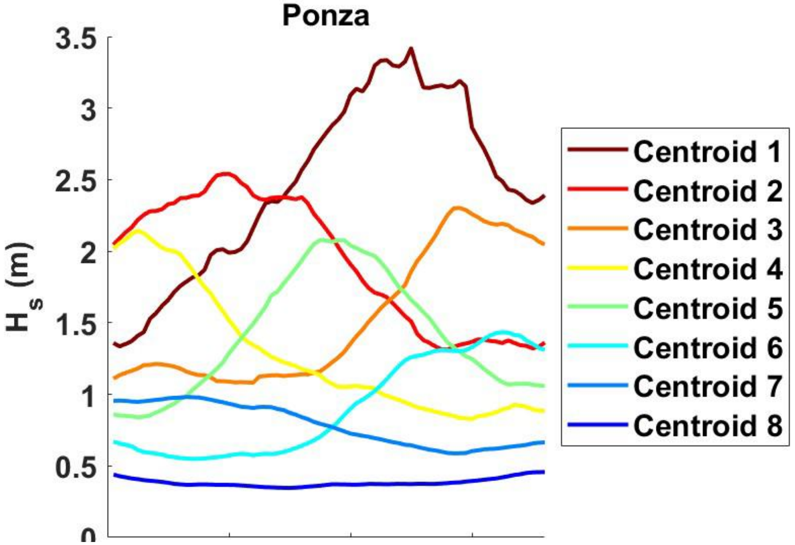
Numerical Results

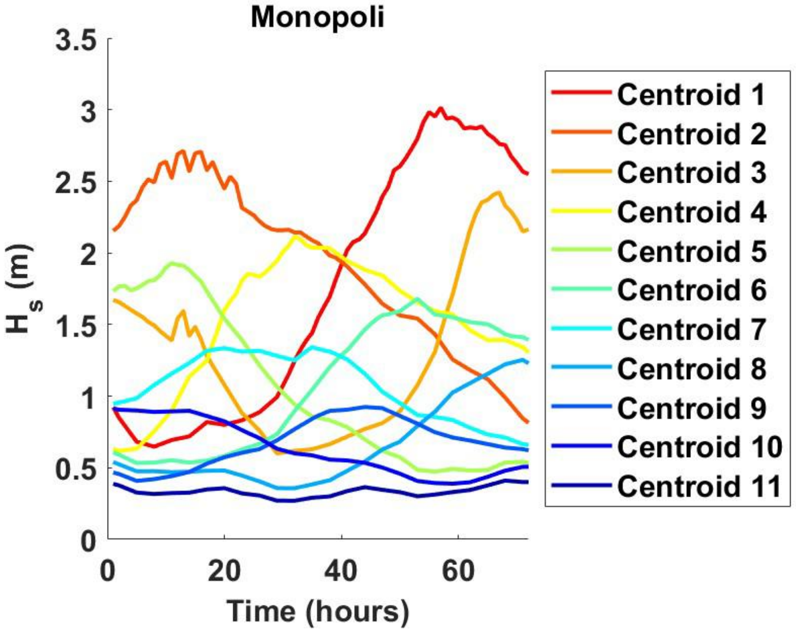
Buoy	Max Wave Height	Optimal Clusters	Original Data Samples	Reduced Data Samples
Mazara	6m	9	81k	12k
Ponza	3.5m	8	96k	14k
Monopoli	3m	11	95k	14k
Ancona	3.5m	9	131k	19k

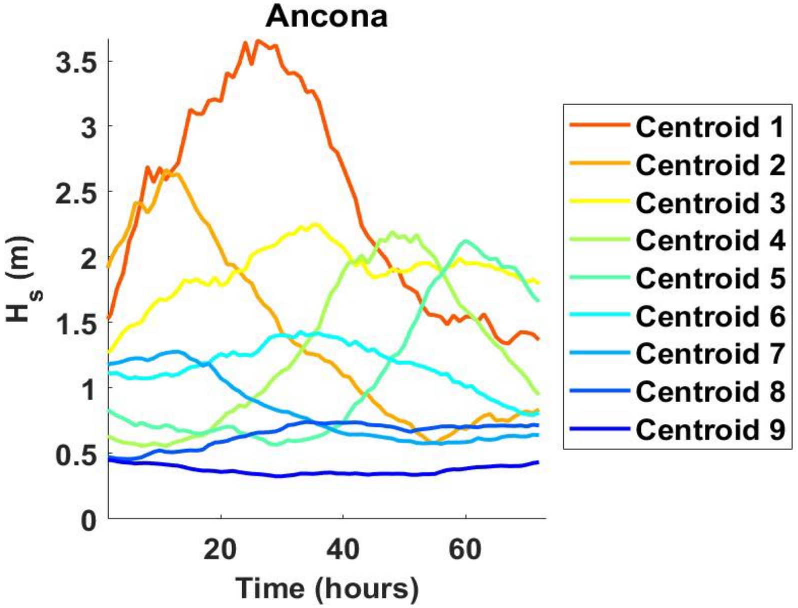


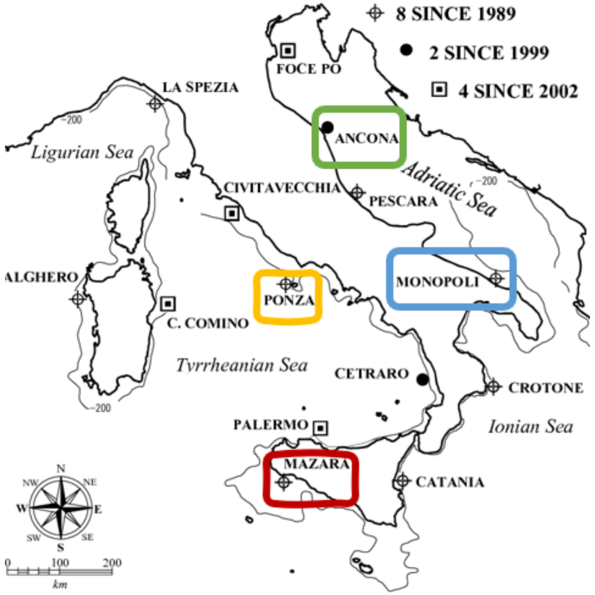
Cross-Location Results


Mazara del Vallo

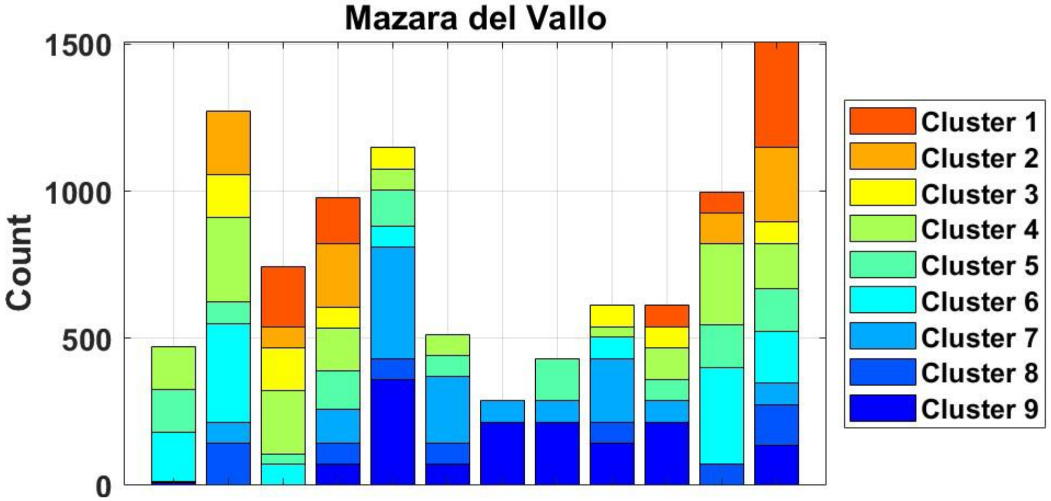

Ponza

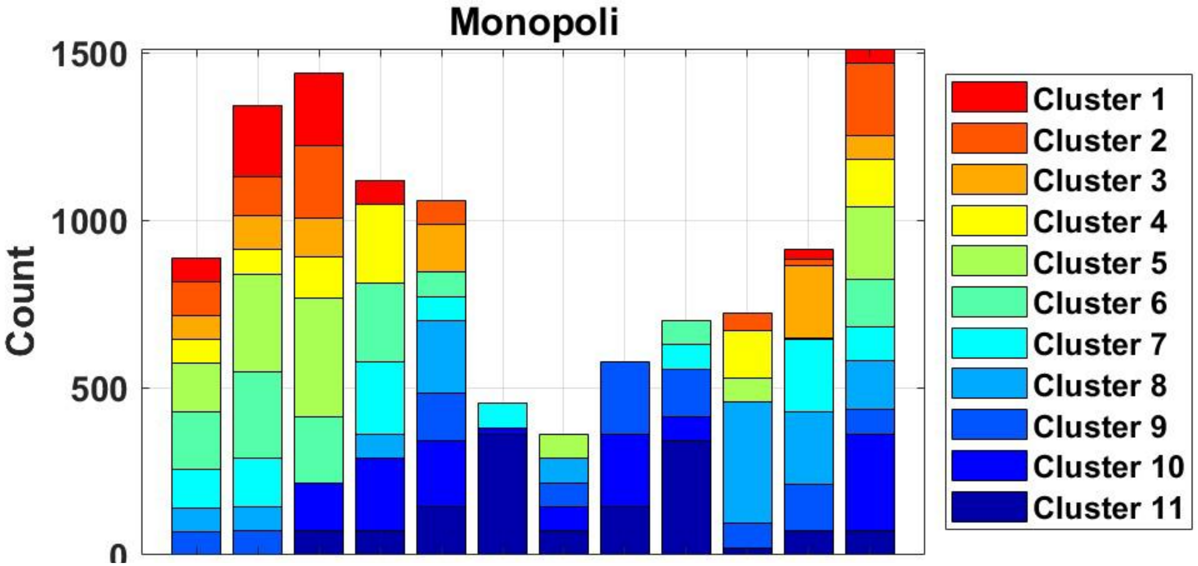

Monopoli

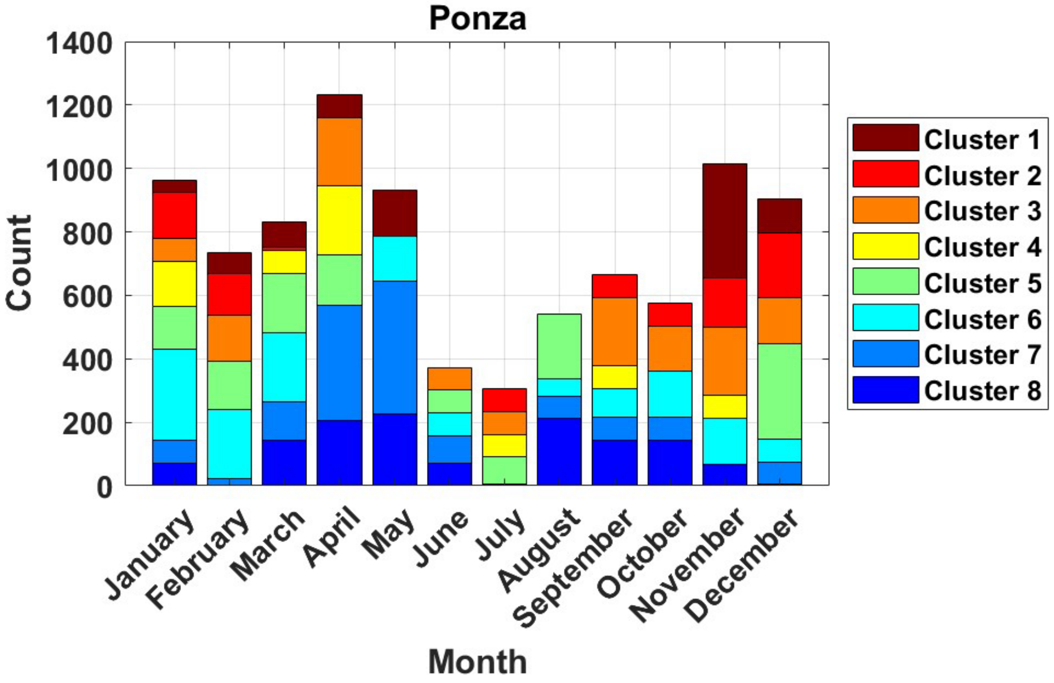

Ancona

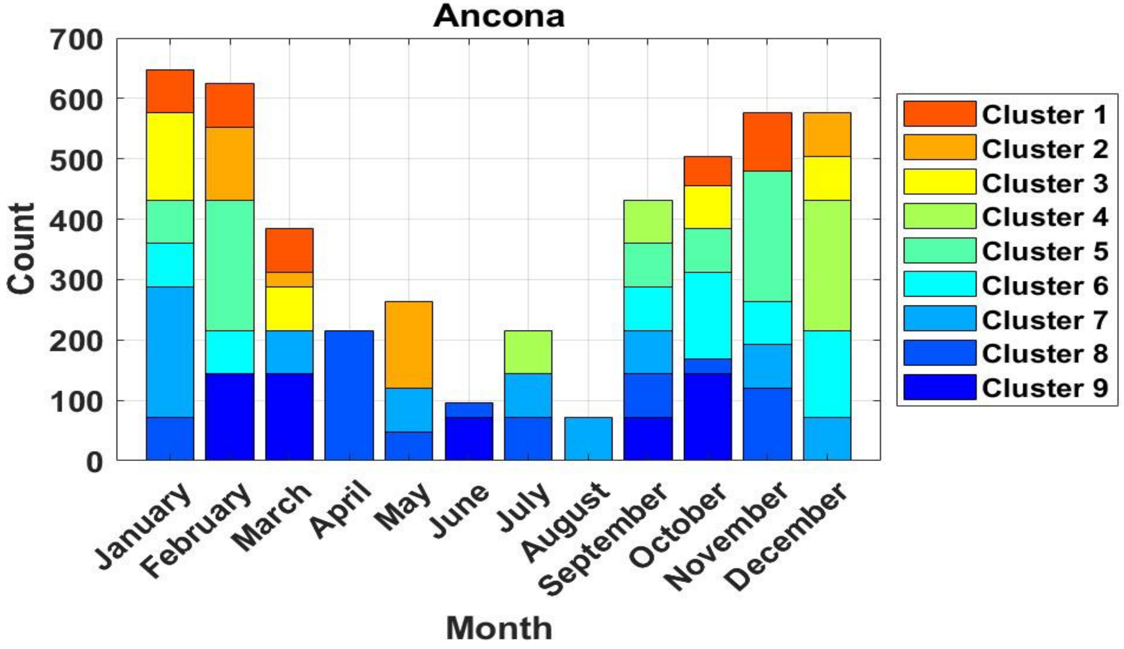


Cross-Location Results









Prediction Analisys

Test	MSE	RMSE	R²
Random test 1	0.327166	0.571984	0.749032
Random test 2	0.161600	0.401996	0.876037
Random test 3	0.312296	0.558834	0.760439
Random test 4	0.189582	0.435410	0.854573
Random test 5	0.282840	0.531827	0.783035
Random test 6	0.268251	0.517930	0.794226
Random test 7	0.246536	0.496524	0.810883
Random test 8	0.217890	0.466786	0.832858
Random test 9	0.334638	0.578479	0.743301
Random test 10	0.297521	0.545455	0.771773
Clustering-based test	0.143131	0.378324	0.890206

Validation:

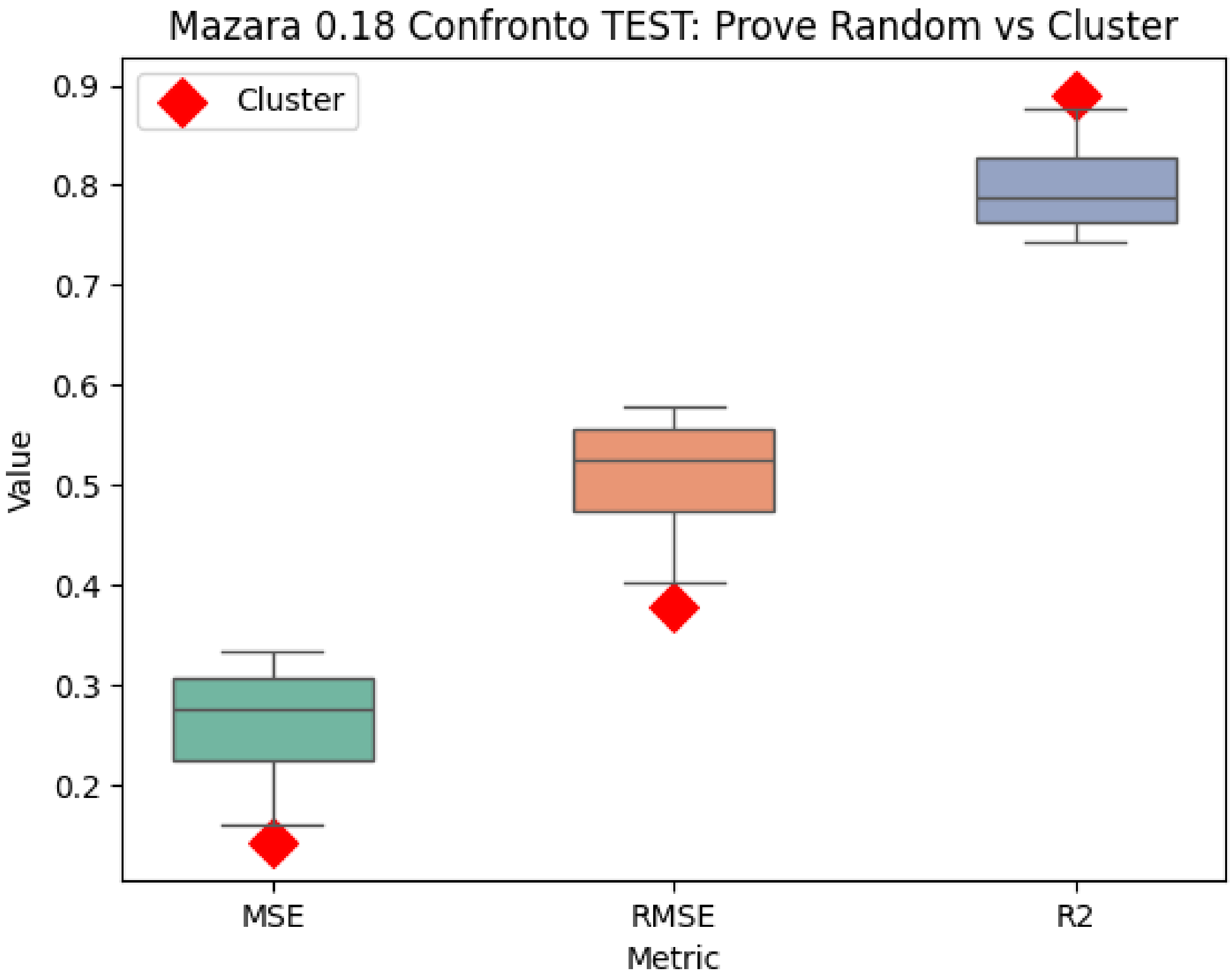
- DNN: CNN + LSTM

Buoy Dataset:

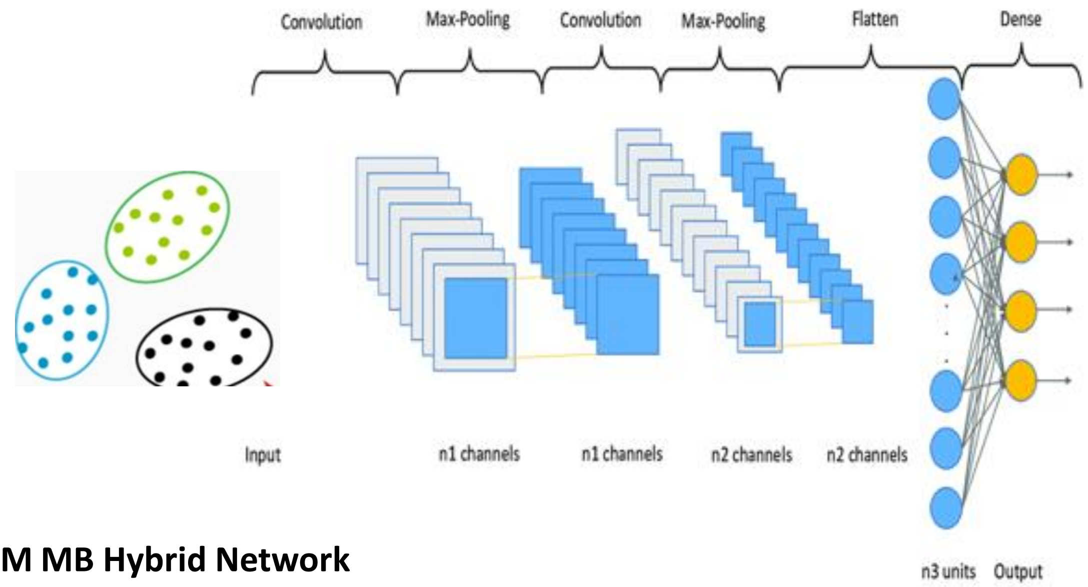
- Mazara del Vallo

L. Patanè, C. Iuppa, C. Faraci, and M. G. Xibilia, “A deep hybrid network for significant wave height estimation,” Ocean Modelling, vol. 189, p. 102363, 2024.

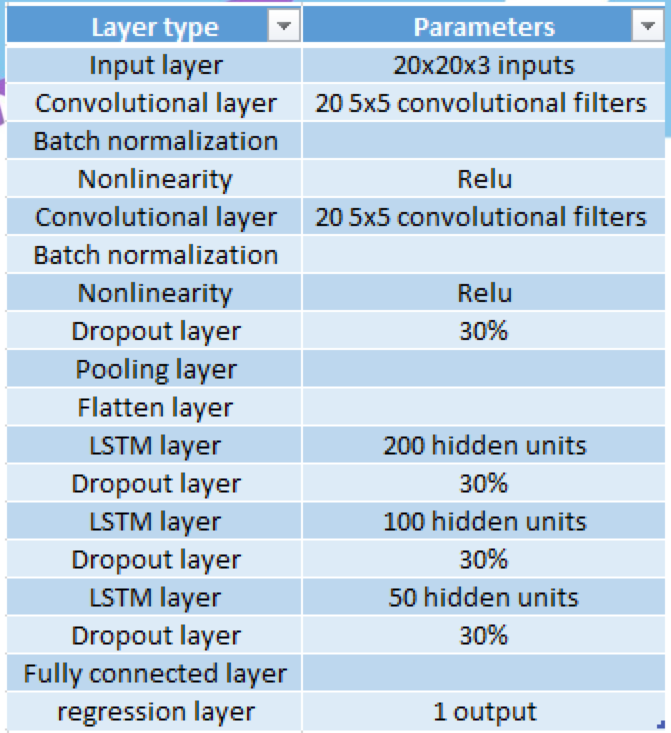
Prediction Analisys



Prediction Analisys



CNN-LSTM MB Hybrid Network

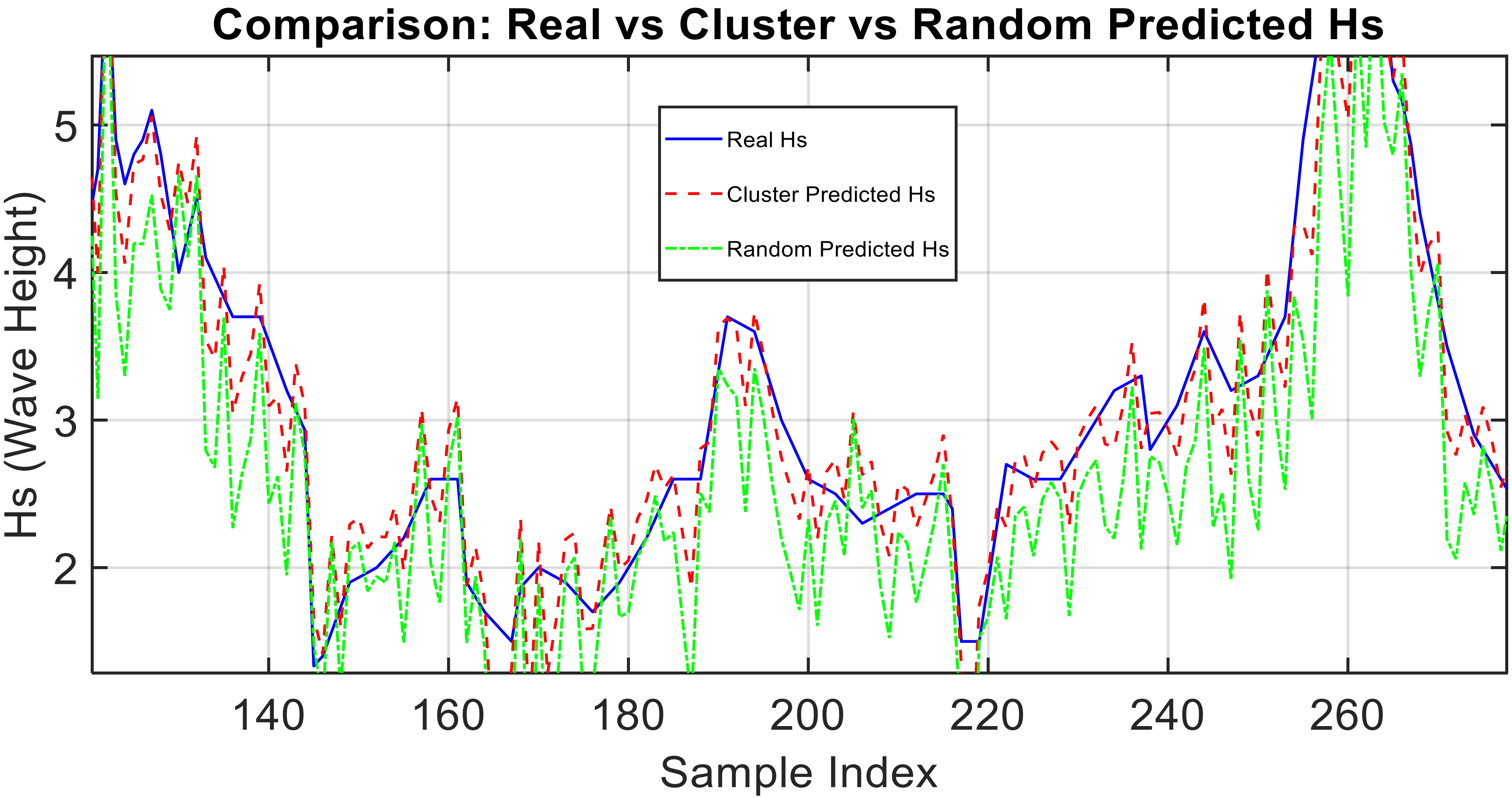


Layer type	Parameters
Input layer	20x20x3 inputs
Convolutional layer	20 5x5 convolutional filters
Batch normalization	
Nonlinearity	Relu
Convolutional layer	20 5x5 convolutional filters
Batch normalization	
Nonlinearity	Relu
Dropout layer	30%
Pooling layer	
Flatten layer	
LSTM layer	200 hidden units
Dropout layer	30%
LSTM layer	100 hidden units
Dropout layer	30%
LSTM layer	50 hidden units
Dropout layer	30%
Fully connected layer	
regression layer	1 output

Test	MSE	RMSE	R^2
Random -based test	0.26383	0.51052	0.79762
Clustering-based test	0.143131	0.378324	0.890206

L. Patanè, C. Iuppa, C. Faraci, and M. G. Xibilia, “A deep hybrid network for significant wave height estimation,” Ocean Modelling, vol. 189, p. 102363, 2024.

Real VS Predicted



18 dicembre 2025

Conclusions and future work

- This work introduces a **clustering-based methodology for data reduction** in the modeling of significant wave height across the Mediterranean Sea.

- By segmenting the original time series into fixed-duration windows and applying K-means clustering, we identify **representative subsets of the wave dynamics with minimal information loss.**

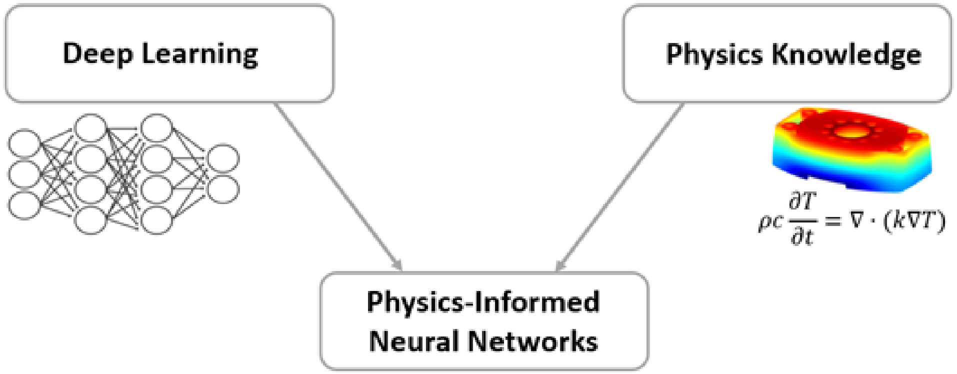
- Applied to data from four buoys, **Mazara del Vallo, Ponza, Monopoli, and Ancona**, the method successfully captures key marine phenomena such as storm events and seasonal trends.

- **Cluster centroids exhibit interpretable characteristics** and provide a compact yet informative basis for predictive model training.

- Overall, the proposed strategy facilitates both **data efficiency** and **model generalization**, offering a scalable solution for wave forecasting in large and heterogeneous maritime domains.

Conclusions and future work

- Extension of the Dataset
- **PINN (Physics Informed Neural Network)**

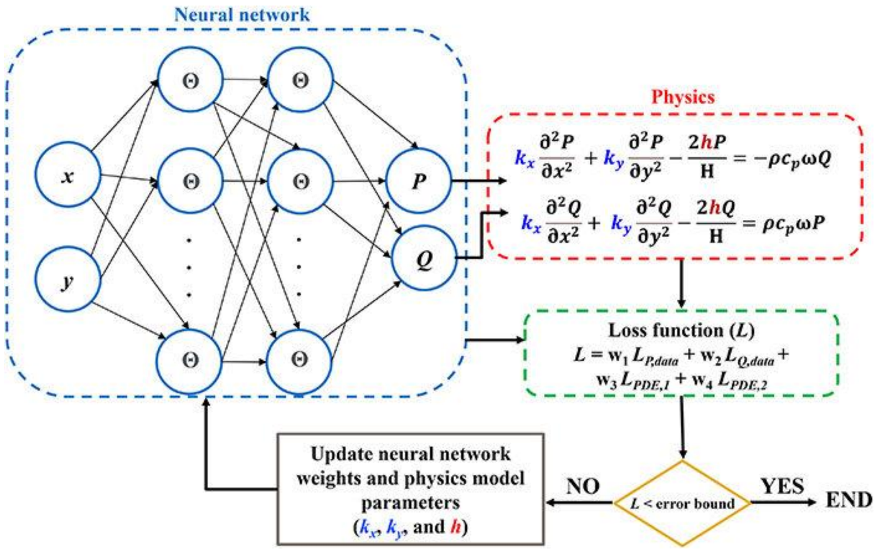


Deep Learning

Physics Knowledge

$$\rho c \frac{\partial T}{\partial t} = \nabla \cdot (k \nabla T)$$

Physics-Informed Neural Networks



$$k_x \frac{\partial^2 P}{\partial x^2} + k_y \frac{\partial^2 P}{\partial y^2} - \frac{2hP}{H} = -\rho c_p \omega Q$$

$$k_x \frac{\partial^2 Q}{\partial x^2} + k_y \frac{\partial^2 Q}{\partial y^2} - \frac{2hQ}{H} = \rho c_p \omega P$$

Loss function (L)

$$L = w_1 L_{P,data} + w_2 L_{Q,data} + w_3 L_{PDE,1} + w_4 L_{PDE,2}$$

L < error bound

Update neural network weights and physics model parameters (k_x, k_y, and h)

Thanks for the attention