Integrating statistical analysis and data-driven modelling for predicting coastal DO levels near Cyprus

Ekaterini Hadjisolomou[1,2]; Konstantinos Antoniades[3]; Lavrentios Vasiliades[3]; Ioannis Kyriakides[2,4]

[1]Department of Electrical Engineering, Computer Engineering and Informatics, Cyprus University of Technology, Cyprus; [2]Cyprus Marine & Maritime Institute, Cyprus; [3]Department of Fisheries and Marine Research, Ministry of Agriculture, Rural Development and the Environment, Cyprus; [4]Department of Engineering, University of Nicosia, Cyprus

Acknowledgments: This work was co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research and Innovation Foundation (MARI-Sense Project: INTEGRATED/0918/0032; OS Aqua project: INTEGRATED/0918/0046) and the EU H2020 Research and Innovation Programme under GA No. 85758 (CMMI-MaRITeC-X).

Introduction: Motivation

- Hypoxia is one of the most significant environmental problems of the modern world, as it is responsible for the degradation of water quality in many freshwater, coastal, and marine ecosystems.
- Coastal waters are major providers of economical services (e.g., tourist activities, aquaculture) and hypoxia/eutrophication has severe socioeconomic implications that threaten humans' well-being, particularly in areas of the world economically dependent from tourism, such as Cyprus.
- In this presentation we utilized data-driven modelling based on Artificial Neural Networks (ANNs) integrated with statistical analysis (PCA) to predict coastal DO levels and to examine the impact of the associated water quality parameters on hypoxia.

Methodology: Data and Study area

Study area

- The island of Cyprus is located in the Levantine Basin - Eastern Mediterranean, which is one of the most oligotrophic seas in the world, characterized by very low nutrient availability and hence very low primary production (Tselepides et al., 2000)
- Levant's Sea has high temperatures ranging yearly from 16 °C in the winter and up to 26 °C in the summer period.
- Cyprus coastal waters are classified as ultra-oligotrophic (2008/915/EC)
- Occasionally eutrophication events are observed in coastal areas of Cyprus (e.g., in the summers of 1990-1991, 1998, 2004-2005 (UNEP, 2007))

Data

- Environmental parameters from coastal stations with different anthropogenic activities (aquaculture, nearby industrial units)
- Sporadic samples -no regular time intervals- were measured during 2000-2021 (*n*=1552) .
- Parameters: nitrogen species (NH_4, NO_2, NO_3); ortho-phosphates (PO_4); salinity; dissolved oxygen (DO); pH; electrical conductivity (EC); water temperature (WT); Chl-a.

Methodology: Feed-forward Model Development

➢ **Data pre-processing:**

- Input Parameter grouping based on PCA

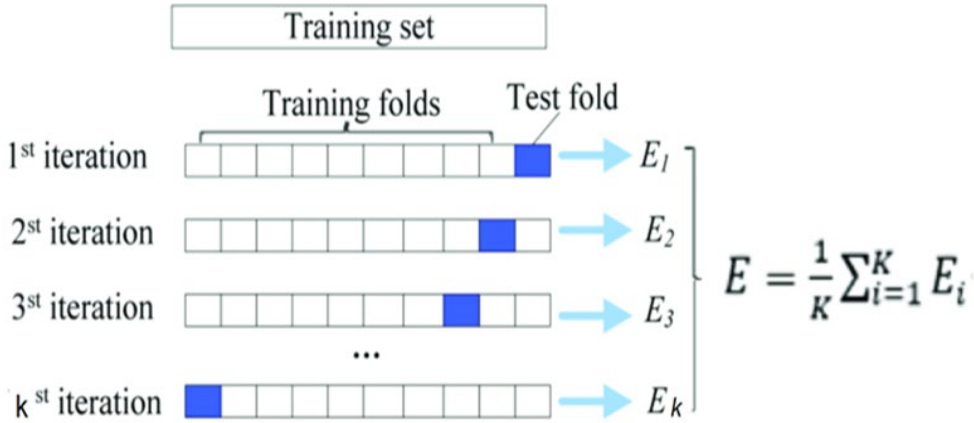
Group	Variable
PC1	EC, salinity, PO4, NO_3
PC2	pH, WT
PC3	Chl-a, NO3, NO2

- Data normalization (Min-Max Scalar)

➢ **ANN's optimal Architecture:**

Inputs	Output	Learning function	Topology
WT, pH, EC, Chl-a, salinity, PO_4, NH_4, NO_2, NO_3	DO	LM algorithm	9-5-1

➢ **Model validation:**

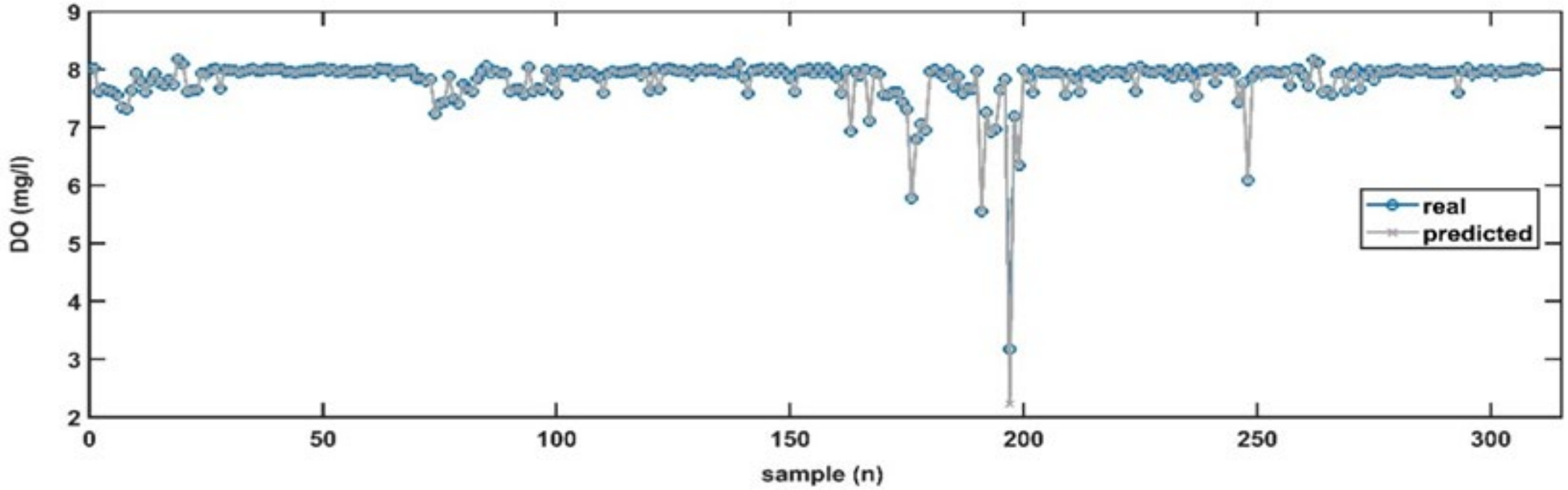


The *k*-fold (*k*=10) cross-validation method gave the best ANN's performance.

➢ **Performance Indices:**

- R^2=0.991
- *RMSE*=0. 0534

Results: real data vs predicted

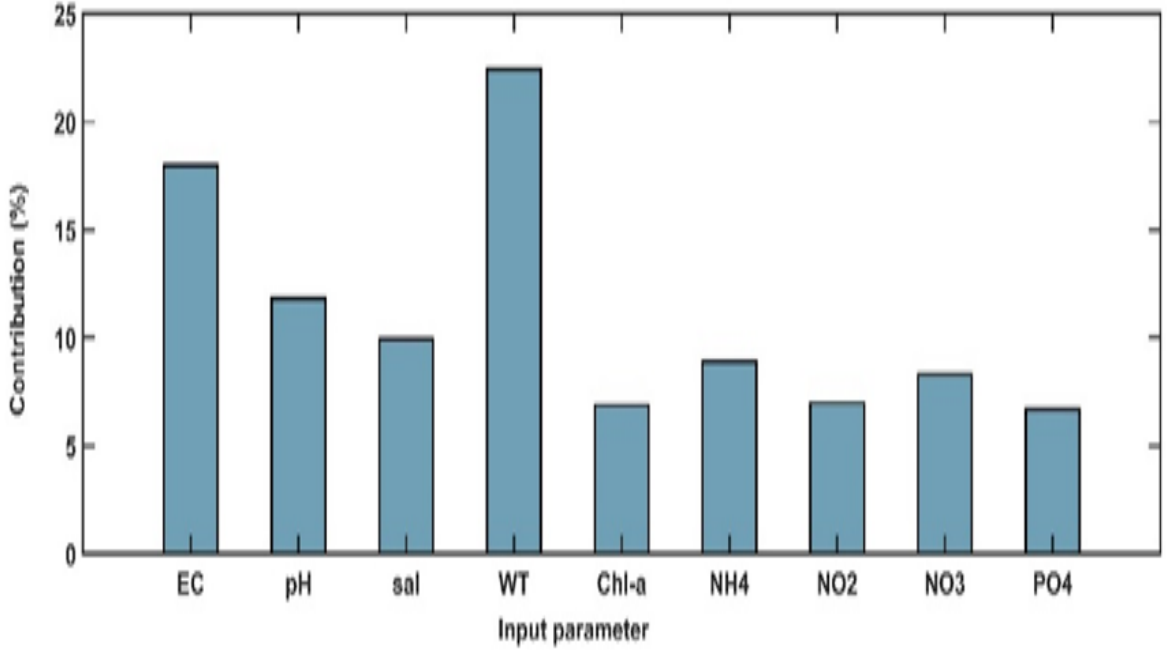


Results: sensitivity analysis

- The “Weights” sensitivity analysis algorithm (or else Garson’s algorithm)
- The contribution of each input parameter to the ANN's output is measured with the relative parameter importance *(I)*

$$I = \frac{\sum_{1=1}^{nT} \sum_{j=1}^{nH} \left|(w_{ij})\right|}{\sum_{k=1}^{nV} \left(\sum_{i=1}^{nT} \sum_{j=1}^{nH} \left|(w_{ij})\right|\right)_k}$$

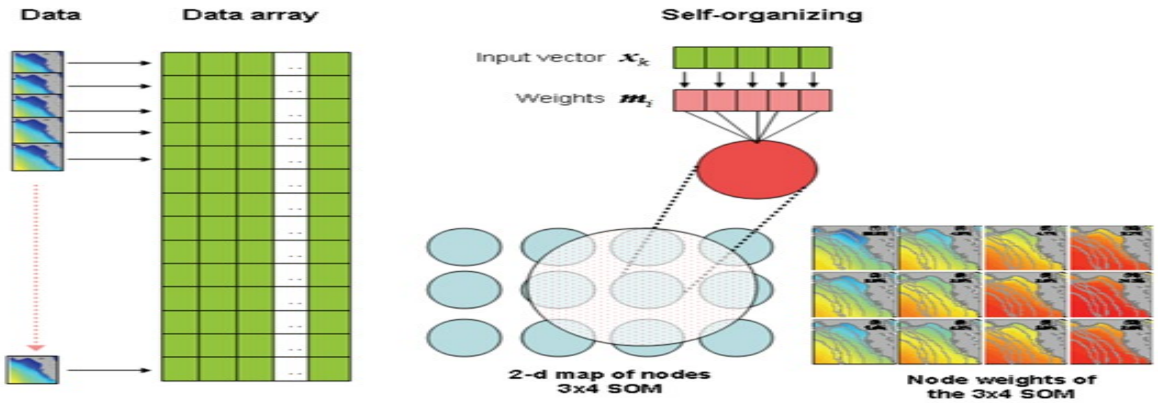
(where nT is the number of time lags, nH the number of hidden neurons and nV is the number of input parameters)



Methodology: Self-organizing map model development

➢ **Data pre-processing:**

Data normalization (log transform)



➢ **ANN's Architecture:**

Inputs	Topology
WT, pH, EC, DO, salinity, PO_4, NH_4, NO_2, NO_3, Chl-a	20x20

SOM algorithm:

1. Initialize the weights with random values.
2. Use of a distance measure to find the best-matching unit (BMU).
3. Update the weights based on the BMU.

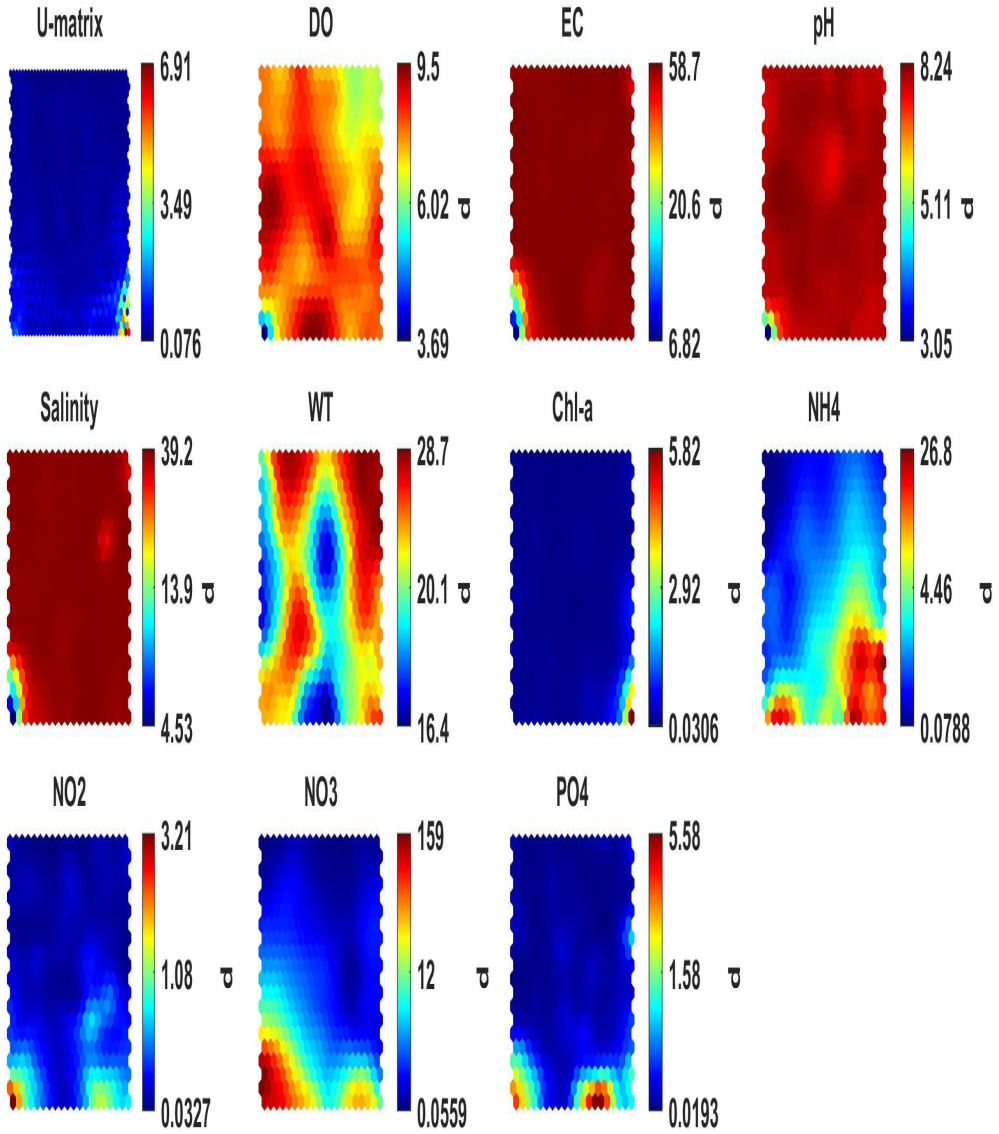
(*analytical presentation is found in the studies of Kangas and Kohonen, 1996; Kohonen, 2001)

Results: PCA results

Variable	PC			
	PC1	PC2	PC3	PC4
DO		-0.5606	0.5492	
EC	-0.4813			
pH		0.3189	0.6795	0.3412
salinity	-0.4840			
WT		0.7277		
Chl-a				0.7958
NH_4				
NO_2				0.3657
NO_3	0.5196			
PO_4	0.3537			
Eigenvalue	3.1336	1.4803	1.1649	1.0304
Percentage of total variance	31.34	14.80	11.65	10.30
Cumulative percentage of variance	31.34	46.14	57.79	68.09

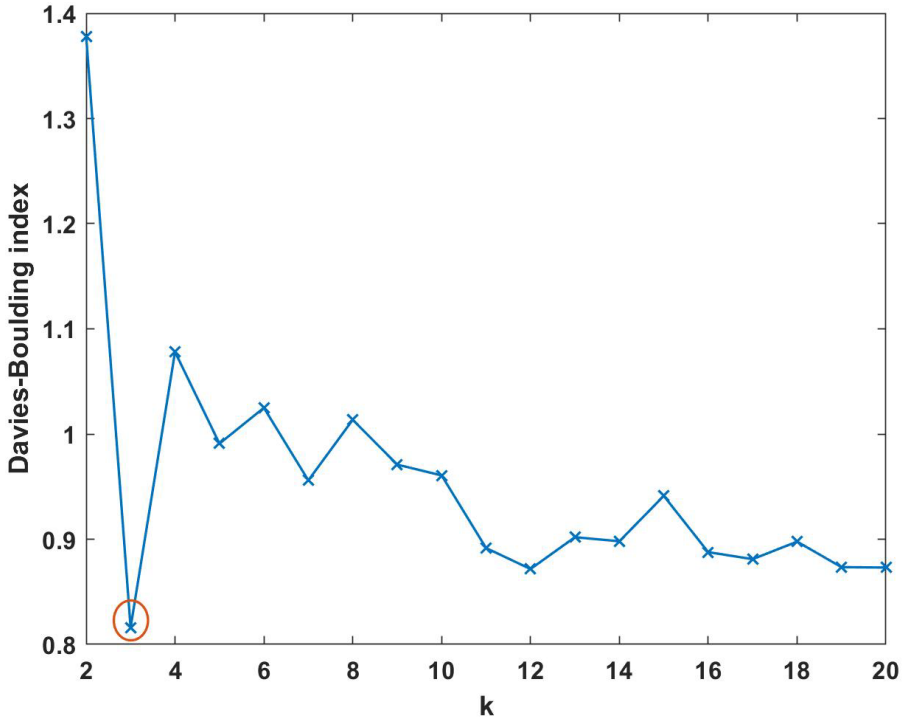
Results: SOM's Component Planes

- EC (+), pH (+) , Salinity(+) : strongly correlated
- DO (+) , EC(+) : partially correlated
- NO2 (+), PO4(+): correlated
- Chl-a (+), WT (+), NH4 (+): partially correlated
- DO (-), NO2 (+), PO4 (+) : partially correlated

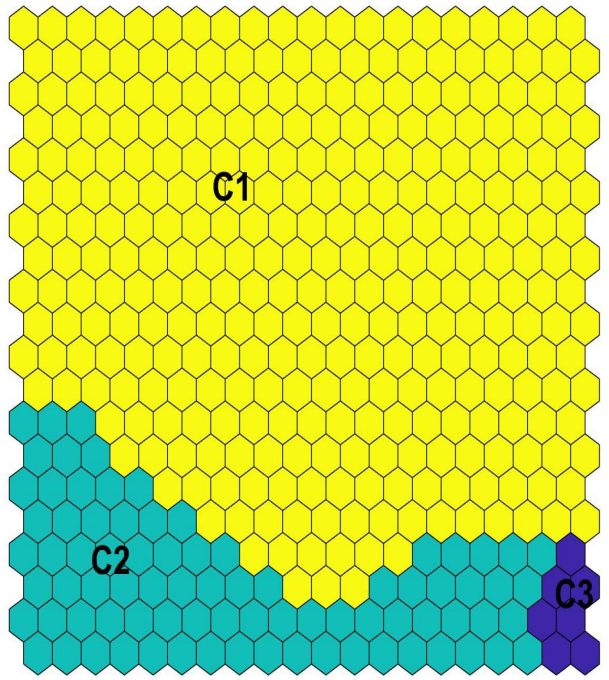


Results: SOM Clustering

The optimal number of clusters minimizes the Davies–Bouldin index when the SOM is implemented by the K-means algorithm



Clustering of the SOM based on the K-means algorithm



Results: Clusters main characteristics

- C1: low Chl-a; high DO; low NO_3, low NH_4; low NO_2; low PO_4
- C2: low Chl-a; high NH_4; high NO_3; high NO_2; high PO_4
- C3: high Chl-a; high NH_4; elevated WT; elevated PO_4; elevated NO_2; elevated NO_3

Conclusions

- The model not only managed to predict with very high accuracy DO, but also managed to capture the environmental mechanisms related with the hypoxia process.
- The sensitivity analysis results revealed that the global warming is greatly affecting the coastal DO levels of Cyprus.
- The increases of nutrients (nitrogen and phosphorus), are negatively impacting (but in a lesser extent than WT) the water quality and are promoting and hypoxia and cultural eutrophication .
- Considering the ANN's good generalization ability, even for extreme / low DO values, several management scenarios could be examined.
- It is proposed that the current feed-forward ANN should be enhanced and recalibrated with denser measurements coming from other available sources (e.g., satellite data), as well as additional variables (e.g., rainfall)

Thank you for your attention!