Modelling of environmental parameters in lake environments using neural networks

Ph. D. Dissertation

Ekaterini Hadjisolomou
Mathematician, MSc Environmental Sciences
Πανεπιστήμιο Πατρών, [Τμήμα Γεωλογίας]

[Χατζησολωμού Αικατερίνη]

© [2019] - Με την επιφύλαξη παντός δικαιώματος
Modelling of environmental parameters in lake environments using neural networks

Ekaterini Hadjisalomou
Ph.D. Dissertation

Advisory Committee:
Prof. George Papatheodorou (Supervisor), Geology Department, University of Patras.
Prof. Nikolaos Lambrakis, Geology Department, University of Patras.
Prof. Evanthia Papastergiadou, Biology Department, University of Patras.

Examination Committee:
Em. Prof. George Ferentinos, Geology Department, University of Patras.
Dr. Nikolaos Kampanis, Research Director, Institute of Applied and Computational Mathematics, Foundation for Research and Technology-Hellas.
As. Prof. Maria Geraga, Geology Department, University of Patras.
As. Prof. Ierotheos Zacharias, Civil Engineering Department, University of Patras.

February 2019
“Only cabbages have no nerves, no worries. And what do they get out of their perfect well-being?”

Karl Gustav Jacob Jacobi (1804-1851), mathematician
Acknowledgments

I would like to express my gratitude to many persons for their support and guidance that I received, in order to achieve this long journey of completing this PhD dissertation. First and foremost, I would like to thank my supervisor Prof. George Papatheodorou for all the support and guidance he gave me. I am proud and honored that Prof. George Papatheodorou is the supervisor of my PhD dissertation, since his scientific achievements can be characterized at least impressive. Thank you Prof. Papatheodorou!!

I would like to deeply thank Prof. Eva Papastergiadou for generously providing her guidance and useful suggestions, which helped me to substantially improve the quality of my work. I gratefully acknowledge the guidance and support of Prof. Nikolaos Lambrakis and his contribution for the fulfillment of my PhD dissertation. Also, I would like to thank Prof. George Ferentinos, Assist. Prof. Maria Geraga, Dr Nikos Kampanis (IACM) and Assist. Prof. Ierotheos Zacharias.

I would like to thank Prof. Eva Papastergiadou (again!) and Dr Kostas Stefanidis for sharing with me the limnological data, based on which the “core” of this PhD thesis was developed. I am glad that I had the chance to cooperate with Dr Kostas Stefanidis, since his contribution to my simulated data implementation was really useful.

Special thanks are devoted to my parents (Diamanto and Christos) and my brother Prokopis for their support, patience and love. Also, special thanks belong to my husband Stratos, for his love and support.
Dedication

To F. and M. for making me smile!
Abstract

Aquatic pollution is one of the most serious environmental issues that the human kind has to deal, especially the last decades. Therefore, the aquatic environments are suffering a great environmental pressure as well. Since the lakes are having about the 90% of the Earth’s liquid freshwater, their proper management is a crucial task in order to maintain their good water quality. Eutrophication is one of the major problems that are observed in the limnetic environments. The eutrophic water is related with a series of environmental problems like anoxia, the existence of harmful cyanotoxins etc. Therefore, this study is focused on the assessment of the environmental parameters that are related with the phenomenon of the eutrophication.

The mathematical models are capable to contribute positively to the restoration of the good water quality status of a limnetic environment. The mathematical models can act as a mean to describe several environmental processes with the use of mathematical relationships. Several types of mathematical models have been applied into the area of environmental sciences, including the artificial neural networks. During the last decades the artificial neural networks have been applied successfully into the field of limnology. It has been documented that the artificial neural networks are superior to other modelling techniques (e.g. multiple linear regression model). This is attributed to the fact that the artificial neural networks can simulate with good correlation the complex non-linear relationships that are describing the environmental processes.

Therefore, several categories of artificial neural networks are applied in this study. The artificial neural networks can be divided into two main categories, those with supervised learning and those with non-supervised learning. In Application 1 and Application 2
supervised artificial neural networks are developed aiming to predict the chlorophyll parameter. Afterwards several sensitivity analysis (one way-sensitivity analysis) algorithms are applied in order to evaluate the environmental parameters with the biggest impact on the artificial neural network, which managed to simulate well (produced small error) the chlorophyll parameter. The synergistic effect of the parameters (two-way sensitivity analysis) is calculated with the use of the “PaD2” algorithm.

In Application 3 a self-organizing map (SOM), which is a non-supervised artificial neural network, is presented. Based on that SOM neural network it is possible the limnological data to be clustered and at the same time the interactions among the environmental parameters to be examined. Furthermore, the statistical methods of Principal Components Analysis and Cluster Analysis are applied and their results are compared with the results produced by the SOM. The results of the Principal Components Analysis and Cluster Analysis methods are with agreement with those produced by the SOM, although the SOM is a superior modelling method because of its advanced visualization abilities to assess the parameters interactions.
Published work

The applications mentioned above (Application 1-3), which are presented in the current Ph.D. Dissertation have been published and are presented in the following order:

1. Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G. Papastergiadou, E. Assessing the Contribution of the Environmental Parameters to Eutrophication with the Use of the “PaD” and “PaD2” Methods in a Hypereutrophic Lake. *International Journal of Environmental Research and Public Health* 2016, 13(8):764, DOI: 10.3390/ijerph13080764

2. Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G. Papastergiadou, E. Evaluating the contributing environmental parameters associated with eutrophication in a shallow lake by applying artificial neural networks techniques. *Fresenius Environmental Bulletin* 2017, 26(5):3200-3208

3. Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G. Papastergiadou, E. Assessment of the Eutrophication-Related Environmental Parameters in Two Mediterranean Lakes by Integrating Statistical Techniques and Self-Organizing Maps. *International Journal of Environmental Research and Public Health* 2018, 15(3):547, DOI: 10.3390/ijerph15030547.
**Abstract per Chapter**

**Abstract-Chapter 1**

Mathematical models are having a major role in the area of environmental/ ecological sciences, including the field of limnology. With the use of mathematical models several biotic and abiotic parameters can be used as model’s input parameters, in order to simulate the arithmetical values of an examined parameter. Based on the mathematical models’ simulations associations/results can be derived about the behavior/impact of the environmental parameters (e.g. association of the soluble reactive phosphorus (SRP) values with the chlorophyll levels). Especially for the chlorophyll parameter simulation, it is found that the artificial neural networks are a better simulation technique comparing with other modelling methods. Additionally, it was found that the artificial neural networks are better to find/examine the patterns among several environmental parameters. This is attributed to the fact that the artificial neural networks are not affected by the non-linearity or the heterogeneity of the simulated system, while it must be noted that the majority of the ecological phenomena are characterized by non-linear behavior.
Abstract-Chapter 2

Artificial neural networks are a branch of artificial intelligence and are mimic the learning function of the human brain, while their architecture is resembling the one of the neural system. Artificial neural networks are divided into two main categories: those with supervised learning and those with un-supervised learning. In the field of limnology, the most widely used artificial neural network is the well-known feed-forward multi-layer perceptron (MLP), trained with the back-propagation training algorithm. The one way-sensitivity analysis of the neural network in regard with the environmental parameters that are used as inputs can be examined with the use of several methods. The main of these methods are: a) the “Perturb” method, b) the “weights” method, c) the Partial Derivatives method or else “PaD method”. The synergistic effect of the combined parameters can be examined with the use of two-way sensitivity analysis by applying the “PaD2”, which is based on the “PaD” algorithm.
Abstract—Chapter 3

Artificial neural networks with un-supervised learning are called after their ability to self-learn. The self-organizing map (SOM) is the most well-known neural network with un-supervised learning. A self-organizing map can be visualized with the use of the U-matrix (unified distance matrix) and its component planes. Based on these visualization abilities the limnological data can be clustered, while at the same time the environmental parameters’ interactions can be examined. Because of its advanced visualization abilities, the self-organizing map is considered a superior clustering method, comparing with the classic statistical methods. The resulted clusters (groups) for the prototypes/outputs of the self-organizing map can be calculated with the U-matrix examination, the k-means algorithm or the use of a dendrogram. The impact/importance of each input parameter of the SOM neural network can be calculated with the calculation of the Cluster Structuring Index (CSI).
Abstract-Chapter 4

Several methodologies exist for the task of water quality modelling, including the statistical methods. The most widely used of these statistical methods are the Principal Components Analysis, Factor Analysis, Cluster Analysis and the Multiple Linear Regression. A variety of published literate that is related with the use of these methods is applied in the field of limnology. Some of these studies are also applying neural networks models for the simulation of the limnological data. These comparative studies almost always concluded that the artificial neural networks are having better (with better accuracy) and more meaningful results. The artificial neural networks with the use of sensitivity analysis can examine of the magnitude/impact of the parameters and therefore to interrelate the parameters (considered as “grey box model”), while the self-organizing maps are benefited with their advanced visualization abilities and therefore are considered a superior modeling method comparing e.g. with PCA. Also, it must be noted that the artificial neural networks are not affected by any heterogeneity or non-linearity of the environmental data.
Abstract—Chapter 5

For the shallow and eutrophic lake Pamvotida the chlorophyll values are simulated with the use of a MLP neural network. The following parameters are used as inputs for the artificial neural network: water temperature (WT), Secchi disk depth (SD), electrical conductivity (EC), dissolved oxygen (DO), pH, dissolved inorganic nitrogen (DIN), dissolved reactive phosphorus (SRP), while the output parameter was the chlorophyll. The artificial neural network simulated with good accuracy the chlorophyll levels ($R = 0.88$). The model’s one-way sensitivity analysis was calculated based on the Partial Derivatives method or else “PaD method”. Based on the sensitivity analysis results it was found that the water temperature (WT) and the dissolved reactive phosphorus (SRP) parameters are having the greatest impact on the model, with relative importance values 50% and 17% respectively. The synergistic effect of the limnological parameters was calculated with the use of SRP-WT two-way sensitivity analysis, by applying the “PaD2” algorithm. The combined parameter SRP-WT was calculated to have the greatest impact with a value 22%. The conclusions for this application/study are that the increase of temperature (meaning climatic change) in combination with increased nutrients levels is having a negative impact on the water quality of lake Pamvotida.
**Abstract-Chapter 6**

Time series for the chlorophyll parameter is predicted/simulated in order to calculate the chlorophyll levels of the next month. For that purpose, an artificial neural network is used and specifically the type of Multilayer Feed-forward (MLF), which is trained with the Levenberg-Marquardt algorithm. The sampling period is between the June 2004-August 2005 and the sampling interval is monthly. For the artificial neural network, the following environmental parameters, which are one-month time lagged, are used as inputs: water temperature (WT), Secchi disk depth (SD), electrical conductivity (EC), dissolved oxygen (DO), pH, dissolved inorganic nitrogen (DIN), dissolved reactive phosphorus (SRP). The simulated result (output) is the chlorophyll value of the next month. The artificial neural network’s results are considered reliable ($R=0.96$). The model’s sensitivity analysis is examined with the application of three different methods: a) the “Perturb” method, b) the “weights” method, c) the Partial Derivatives method or else “PaD method”. Based on the calculated results found by these three methods a mean sensitivity analysis index is proposed, which is averaging any possible computational differences that might result by the three sensitivity methods, regarding the parameters’ importance.
Abstract-Chapter 7

The environmental parameters associated with the eutrophication are assessed in two different morphological lakes, the shallow eutrophic lake Mikri Prespa and the deep mesotrophic lake Megali Prespa. The data were monitored seasonally (spring, summer, autumn) during the period 2006-2008 and are concerning only the Greek part of the lakes. The monitored environmental parameters were the water temperature, Secchi disk depth, electric conductivity, dissolved oxygen, pH, dissolved inorganic nitrogen, total phosphorus. For the assessment of the environmental parameters’ correlation/behavior principal component analysis (PCA) method, cluster analysis and a self-organizing map (SOM) were applied. The principal component analysis (PCA) and the cluster analysis results agreed with the results of the SOM model, although the SOM model was proved to be a more effective method because of its advanced visualization abilities to examine the parameters. The k-means algorithm and a dendrogram were applied on the SOM’s prototypes/outputs and the resulted groups were statistically evaluated. The increased nutrients levels were found to be positively correlated with the increased chlorophyll production. The shallow lake Mikri Prespa was found to be more prone to the climatic change and the seasonality effect than the deep lake Megali Prespa.
Περίληψη

Η υδατική μόλυνση είναι ένα από τα μεγαλύτερα περιβαλλοντικά προβλήματα που έχει να αντιμετωπίσει η ανθρωπότητα ειδικά τις τελευταίες δεκαετίες. Συνεπώς τα λιμναία περιβάλλοντα υφίστανται και αυτά τεράστιες περιβαλλοντικές επιπτώσεις. Δεδομένου ότι οι λίμνες καλύπτουν περίπου το 90% του επιφανειακού υγρού γλυκού νερού της Γης, η ορθή διαχείριση τους είναι πλέον επιτακτική ανάγκη, έτσι ώστε να υπάρχει καλή ποιότητα των υδάτων. Ο ευτροφισμός είναι ένα από τα κυριότερα προβλήματα που παρατηρούνται στα λιμναία περιβάλλοντα. Τα ευτροφικά νερά είναι άμεσα συνδεδεμένα με πληθώρα περιβαλλοντικών προβλημάτων όπως η ανοξία, η συγκέντρωση επικίνδυνων κυανοτοξινών κλπ. Οπότε η παρούσα μελέτη επικεντρώνεται στην εξέταση την περιβαλλοντικών παραμέτρων και την επίδρασή/συσχέτιση τους με το φαινόμενο του ευτροφισμού.

Τα μαθηματικά μοντέλα (πρότυπα) δύνανται να συνεισφέρουν στην προσπάθεια για αποκατάσταση της καλής ποιότητας του νερού σε ένα λιμναίο περιβάλλον. Τα μαθηματικά μοντέλα διαδραματίζουν το ρόλο ενός διάμεσου για την περιγραφή διάφορων περιβαλλοντικών διεργασιών μέσω μαθηματικών σχέσεων. Διάφοροι τύποι μαθηματικών μοντέλων έχουν εφαρμοστεί στις περιβαλλοντικές επιστήμες, ανάμεσα τους και τα τεχνητά νευρωνικά δίκτυα. Τις τελευταίες δεκαετίες έχει γίνει εφαρμογή των τεχνητών νευρωνικών δικτύων με ιδιαίτερη επιτυχία στην λιμνολογία. Τα τεχνητά νευρωνικά δίκτυα έχει παρατηρηθεί ότι υπερτερούν έναντι πολλών άλλων κατηγοριών μαθηματικών προτύπων (π.χ. μοντέλο γραμμικής παλινδρόμησης). Αυτό έγκειται στο γεγονός ότι μπορούν να προσομοιώσουν με καλή ακρίβεια τις πολύπλοκες μη γραμμικές σχέσεις που συνήθως περιγράφουν τις περιβαλλοντικές διαδικασίες.

Οπότε διάφορες κατηγορίες τεχνητών νευρωνικών δικτύων εφαρμόζονται στην παρούσα μελέτη. Τα τεχνητά νευρωνικά δίκτυα μπορούν να χωριστούν σε δύο κύριες κατηγορίες,
αυτά με επιβλεπόμενη μάθηση και αυτά με μη επιβλεπόμενη μάθηση. Στις Εφαρμογές 1 και Εφαρμογή 2 υλοποιούνται τεχνητά νευρωνικά δίκτυα με επιβλεπόμενη μάθηση, με στόχο να προβλέψουν την παράμετρο της χλωροφύλλη. Ενώ μετέπειτα διάφοροι αλγόριθμοι ανάλυσης ευαισθησίας (one way-sensitivity analysis) εφαρμόζονται στο τεχνητό νευρωνικό δίκτυο, έτσι ώστε να εντοπιστούν οι περιβαλλοντικές παράμετροι με την πιο μεγάλη επίδραση στο μοντέλο, το οποίο προσομοιώνει επιτυχώς (μικρό σφάλμα) την παραγωγή χλωροφύλλης. Η συνεργαστική δράση (two-way sensitivity analysis) των συζευγμένων περιβαλλοντικών παραμέτρων υπολογίζεται με την χρήση του “PaD2” αλγορίθμου.

Στην Εφαρμογή 3 υλοποιείται ένα τεχνητό νευρωνικό δίκτυο με μη επιβλεπόμενη μάθηση του τύπου των αυτό-οργανωτικών χαρτών (self-organizing map (SOM)), βάση του οποίου μπορεί να γίνει διαχωρισμός σε κλάσεις των λιμνολογικών δεδομένων, ενώ ταυτόχρονα γίνεται μελέτη της αλληλεπίδρασης μεταξύ των περιβαλλοντικών παραμέτρων. Παράλληλα οι στατιστικές μέθοδοι της Ανάλυσης Κυρίων Συνιστώσων (Principal Components Analysis) και της Ανάλυσης Συστάδων (Cluster Analysis) εφαρμόζονται και τα αποτελέσματα τους συγκρίνονται με αυτά του αυτό-οργανωτικού χάρτη.

Τα αποτελέσματα της ανάλυσης κύριων συνιστώσων και της ανάλυσης κατά συστάδες βρίσκονται να είναι σε συμφωνία με τα αποτελέσματα του αυτό-οργανωτικού χάρτη, αν και ο αυτό-οργανωτικός χάρτης βρέθηκε να είναι πιο αποτελεσματικός μέθοδος χάρη στις προηγμένες απεικονιστικές ικανότητες που έχει για να συσχετίζει παραμέτρους.
Περίληψη ανά Κεφάλαιο

Περίληψη-Κεφάλαιο 1

Ο ρόλος των μαθηματικών μοντέλων (πρότυπων) είναι ιδιαίτερα σημαντικός στις περιβαλλοντικές/οικολογικές επιστήμες, συμπεριλαμβανομένης και της λιμνολογίας. Με τη χρήση μαθηματικών προτύπων διάφορες βιοτικές και αβιοτικές παράμετροι μπορούν να χρησιμοποιηθούν σαν παράμετροι εισόδου étσi ώστε να προσομοιώσουν τις τιμές μιας ζητούμενης παραμέτρου. Βάση των προβλέψεων των μαθηματικών προτύπων παράγονται συσχετισμοί/αποτελέσματα σχετικά με την συμπεριφορά/επίδραση των περιβαλλοντικών παραμέτρων (π.χ. συσχέτιση των τιμών του διαλυτού ενεργού φώσφορου (SRP) με τα επίπεδα της χλωροφύλλης). Τα τεχνητά νευρωνικά δίκτυα έχει βρεθεί ότι είναι καλύτερης μέθοδος υπολογισμού, ειδικά όσο αφορά τον υπολογισμό της παραμέτρου της χλωροφύλλης, σε σχέση με άλλες μεθόδους μοντέλοποιησης. Παράλληλα έχει βρεθεί ότι τα τεχνητά νευρωνικά δίκτυα είναι πιο καλά στην εύρεση/μελέτη των μοτίβων ανάμεσα σε διάφορες περιβαλλοντικές παραμέτρους. Αυτό έγκειται στο γεγονός ότι τα τεχνητά νευρωνικά δίκτυα δεν επηρεάζονται από την μη γραμμικότητα ή την ετερογένεια του συστήματος που προσομοιώνουν, ενώ ταυτόχρονα πρέπει να λάβουμε υπόψη το γεγονός ότι τα πιο πολλά οικολογικά φαινόμενα χαρακτηρίζονται από μη γραμμική συμπεριφορά.
Περίληψη-Κεφάλαιο 2

Τα τεχνητά νευρικά δίκτυα είναι ένα είδος τεχνητής νοημοσύνης και μιμούνται την λειτουργία εκμάθησης του ανθρώπινου εγκέφαλου, ενώ η αρχιτεκτονική τους προερχόμενη, τον περιβάλλον συστήματος. Τα τεχνητά νευρικά δίκτυα χωρίζονται σε δύο κατηγορίες: αυτά με επιβλεπόμενη μάθηση και αυτά με μη επιβλεπόμενη μάθηση. Το πιο ευρέως χρησιμοποιούμενο τεχνητό νευρικό δίκτυο στον τομέα της λιμνολογίας είναι το πολύ γνωστό με διάδοση "προς τα εμπρός" (feed-forward) πολυστρωματικό δίκτυο αισθητήρα (Multi Layer Perceptron (MLP)) εκπαιδευόμενο με τον αλγόριθμο οπισθοδιάδοσης (back-propagation training algorithm). Η μονόδρομη ανάλυση ευαισθησίας (one-way-sensitivity analysis) για τις διάφορες περιβαλλοντικές παραμέτρους που χρησιμοποιούνται ως εισαγόμενες μεταβλητές για το τεχνητό νευρικό δίκτυο μπορεί να εξεταστεί βάση διάφορων μεθόδων, με τις κυρίωτερες από αυτές να είναι: α) η μέθοδος “Perturb”, β) η μέθοδος των βαρών (weights method), γ) η μέθοδος των μερικών διαφορικών (Partial Derivatives method ή "PaD method"). Η συνεργαστική δράση των συζευγμένων περιβαλλοντικών παραμέτρων μπορεί να εξεταστεί με την αμφίδρομη ανάλυση ευαισθησίας (two-way sensitivity analysis) κάνοντας χρήση του “PaD2” αλγοριθμού, ο οποίος βασίζεται πάνω στον "PaD" αλγόριθμο.
Περίληψη-Κεφάλαιο 3

Τα τεχνητά νευρονικά δίκτυα με μη επιβλεπόμενη μάθηση έχουν ονομαστεί έτσι εξαιτίας της ιδιότητας τους να μπορούν να αυτό-εκπαιδεύονται. Ο αυτό-οργανωτικός χάρτης (self-organizing map (SOM)) είναι το πιο γνωστό τεχνητό νευρονικό δίκτυο με μη επιβλεπόμενη μάθηση. Ένας αυτό-οργανωτικός χάρτης μπορεί να απεικονιστεί με τη χρήση της U-matrix (unified distance matrix) και των συστατικών επιπέδων (component planes). Βάση αυτών των απεικονιστικών του δυνατοτήτων μπορεί να γίνει διαχωρισμός σε κλάσεις των λιμνολογικών δεδομένων, ενώ ταυτόχρονα γίνεται μελέτη της αλληλεπίδρασης μεταξύ των περιβαλλοντικών παραμέτρων. Χάρη σε αυτές τις προηγμένες απεικονιστικές ικανότητες που έχει για να συσχετίζει παραμέτρους, ένας αυτό-οργανωτικός χάρτης θεωρείται πιο αποτελεσματική μέθοδος για ομαδοποίηση συγκριτικά με τις κλασσικές στατιστικές μεθόδους. Οι κλάσεις (ομάδες) που προκύπτουν για τα πρωτότυπα (prototypes)/αποτελέσματα του αυτό-οργανωτικού χάρτη μπορούν να υπολογιστούν είτε με την μελέτη της U-matrix, είτε με τη χρήση του k-means όλγορθμου και τέλος με τη χρήση ενός δενδρογράμματος. Η βαρύτητα/δράση που έχει κάθε παράμετρος εισόδου στο SOM νευρονικό υπολογίζεται με την εύρεση του δείκτη Cluster Structuring Index (CSI).
Περίληψη-Κεφάλαιο 4

Για την μοντελοποίηση της ποιότητας του νερού υπάρχουν διάφορες μεθοδολογίες, συμπεριλαμβανομένων και των στατιστικών μεθόδων. Οι πιο ευρέως εφαρμοσμένες από τις στατιστικές μεθόδους είναι η Ανάλυση Κύριων Συνιστωσών (Principal Components Analysis), η Παραγοντική Ανάλυση (Factor Analysis), η Ανάλυση Συστάδων (Cluster Analysis) και η Πολλαπλή Γραμμική Παλινδρόμηση (Multiple Linear Regression). Πληθώρα δημοσιευμένων εργασιών με τις πιο πάνω μεθόδους συναντούνται και στον τομέα της λιμνολογίας. Κάποιες από αυτές τις εργασίες κάνουν ταυτόχρονα και προσομοίωση των λιμνολογικών δεδομένων και με χρήση τεχνητών νευρωνικών δικτύων. Σχεδόν πάντοτε έχει βρεθεί από αυτές τις συγκριτικές εργασίες ότι τα τεχνητά νευρωνικά δίκτυα παράγουν καλύτερα (με μεγαλύτερη ακρίβεια) και πιο επουσιώδη αποτελέσματα. Με τη χρήση της ανάλυσης αυσιτιστής μπορεί να γίνει εξέταση της βαρύτητας/επίδρασης των παραμέτρων και να γίνουν κάποιοι συσχετισμοί ανάμεσα στις παραμέτρους (θεωρούνται "grey box model"), ενώ οι αυτο-οργανωτικοί χάρτες έχουν προηγμένες απεικονιστικές δυνατότητες των αποτελεσμάτων γεγονός που τους καθιστά ως ανώτερη μέθοδο μοντελοποίησης σε σχέση π.χ. με την PCA. Επίσης πρέπει να σημειωθεί ότι τα τεχνητά νευρωνικά δίκτυα δεν επηρεάζονται από ενδεχόμενη μη γραμμικότητα ή επιφορές των περιβαλλοντικών δεδομένων.
Περίληψη-Κεφάλαιο 5

Τα επίπεδα της χλωροφύλλης στη ρηχή και ευτροφική λίμνη Παμβώτιδα προσομοιώνονται με τη χρήση ενός MLP νευρωνικού μοντέλου. Σαν παράμετροι εισόδου για το νευρωνικό δίκτυο χρησιμοποιήθηκαν οι παράμετροι: θερμοκρασία νερού (WT), βάθος δίσκου Secchi (SD), ηλεκτρική αγωγιμότητα (EC), διαλυμένο οξυγόνο (DO), pH, διαλυτό ανόργανο άζωτο (DIN), διαλυτό ενεργό φώσφορο (SRP) και παράμετρος εξόδου (αποτέλεσμα) ήταν η χλωροφύλλη. Το τεχνητό νευρωνικό δίκτυο προσομοίωσε με αξιοπιστία τα επίπεδα της χλωροφύλλης (R= 0.88). Η ανάλυση ευαισθησίας (one-way sensitivity analysis) του μοντέλου υπολογίστηκε βάση του αλγορίθμου των μερικών διαφορικών (Partial Derivatives method ή “PaD method”). Βάση των αποτελεσμάτων της ανάλυσης ευαισθησίας βρέθηκε ότι η παράμετρος που προσμετρά την θερμοκρασία νερού (WT) και η παράμετρος για το διαλυτό ενεργό φώσφορο (SRP) έχουν την πιο μεγάλη επίδραση στο μοντέλο, με τιμές σχετικής βαρύτητας 50% και 17% αντίστοιχα. Η συνεργιστική δράση των λιμνολογικών παραμέτρων υπολογίστηκε με την αμφίδρομη ανάλυση ευαισθησίας (two-way sensitivity analysis) κάνοντας χρήση του “PaD2” αλγορίθμου. Η συζευγμένη παράμετρος SRP-WT βρέθηκε να έχει την πιο μεγάλη σχετική βαρύτητα με τιμή 22%. Τα συμπεράσματα που εξάγονται από την συγκεκριμένη μελέτη/εφαρμογή είναι ότι η αύξηση της θερμοκρασίας (δηλαδή κλιματική αλλαγή) σε συνδυασμό με τις αυξημένες επίπεδα θρεπτικών συμβάλλουν αρνητικά στην ποιότητα του νερού της λίμνης Παμβώτιδας.
Περίληψη-Κεφάλαιο 6

Η χρονοσειρά για την παράμετρο της χλωροφύλλης προβλέπεται/προσομοιώνεται έτσι ώστε να γίνεται υπολογισμός των επιπέδων της χλωροφύλλης για τον επόμενο μήνα. Για τον σκοπό αυτό γίνεται χρήση ενός τεχνητού νευρωνικού δικτύου, και συγκεκριμένα ενός πολυστοιβαδικού δικτύου (Multilayer Feed-forward (MLF)) το οποίο έχει εκπαιδευθεί με τον αλγόριθμο Levenberg–Marquardt. Η περίοδος δειγματοληψίας είναι μεταξύ του διαστήματος Ιουνίου 2004-Αύγουστο 2005 μετρημένα ανά μήνα. Οι περιβαλλοντικές παράμετροι οι οποίες χρησιμοποιούνται σαν δεδομένα εισόδου (inputs) για το νευρωνικό μοντέλο έχουν χρονική υστέρηση (time lagged) ένα μήνα και είναι οι εξής: θερμοκρασία νερού (WT), βάθος δίσκου Secchi (SD), ηλεκτρική αγωγιμότητα (EC), διαλυμένο οξυγόνο (DO), pH, διαλυτό ανόργανο άζωτο (DIN), διαλυτό ενεργό φώσφορο (SRP). Το προσομοιωμένο αποτέλεσμα (output) είναι η τιμή της συγκέντρωσης της χλωροφύλλης για τον επόμενο μήνα. Τα αποτελέσματα του νευρωνικού μοντέλου είναι αξιόπιστα ($R=0.96$). Επίσης η ανάλυση ευαισθησίας του μοντέλου εξετάζεται με την χρήση τριών διαφορετικών μεθόδων: α) η μέθοδος “Perturb”, β) η μέθοδος των βαρών (weights method), γ) η μέθοδος των μερικών διαφορικών (Partial Derivatives method ή “PaD method”). Ένας μέσος δείκτης ευαισθησίας προτείνεται βάση των αποτελεσμάτων που προκύπτουν από τις τρεις πιο πάνω μεθόδους ανάλυσης ευαισθησίας, ο οποίος αντισταθμίζει τυχών υπολογιστικές διαφορές που μπορεί να προκύπτουν για τη σειρά σημαντικότητας των παραμέτρων.
Περίληψη-Κεφάλαιο 7

Οι περιβαλλοντικές παράμετροι που σχετίζονται με το φαινόμενο του ευτροφισμού αξιολογήθηκαν σε δύο διαφορετικές μορφολογικά λίμνες, την ρηχή ευτροφική λίμνη Μικρή Πρέσπα και την βαθιά μεσοτροφική λίμνη Μεγάλη Πρέσπα. Τα δεδομένα έχουν συλλεχθεί εποχικά (άνοιξη, φθινόπωρο, καλοκαίρι) κατά την χρονική περίοδο 2006-2008 και αφορούν το Ελληνικό κομμάτι των λιμνών. Οι περιβαλλοντικές παράμετροι που καταγράφηκαν ήταν: θερμοκρασία του νερού, βάθος δίσκου Secchi, βάθος νερού, ηλεκτρική αγωγιμότητα, διαλυμένο οξυγόνο, pH, διαλυτό ανόργανο άζωτο, ολικός φώσφορος. Για την αξιολόγηση της συσχέτισης/συμπεριφοράς των περιβαλλοντικών παραμέτρων εφαρμόστηκαν η μέθοδος της ανάλυσης κύριων συνιστωσών (PCA), η ανάλυση κατά συστάδες-ομάδες (cluster analysis), και ένα μη επιβλεπόμενο τεχνητό νευρωνικό δίκτυο της κατηγορίας των αυτό-οργανωτικών χαρτών (self-organizing map(SOM)). Τα αποτελέσματα της ανάλυσης κύριων συνιστωσών (PCA) και της ανάλυσης κατά συστάδες (CA) ήταν σε συμφωνία με τα αποτελέσματα του SOM μοντέλου, αν και το SOM μοντέλο είναι πιο αποτελεσματική μέθοδος χάρη στις προηγμένες απεικονιστικές ικανότητες που έχει για να συσχετίζει παραμέτρους. Ο αλγόριθμος k-means καθώς και ένα δενδρόγραμμα εφαρμόστηκαν πάνω στα πρωτότυπα (prototypes)/ αποτελέσματα του αυτό-οργανωτικού χάρτη και οι ομάδες που προέκυψαν μελετήθηκαν στατιστικά. Τα αυξημένα επίπεδα θρεπτικών βρέθηκαν να συσχετίζονται θετικά με την αυξημένη παραγωγή χλωροφύλλης. Επίσης βρέθηκε ότι και η θερμοκρασία διαδραματίζει σημαντικό ρόλο ως προς την αυξημένη παραγωγή χλωροφύλλης. Η δε ρηχή λίμνη Μικρή Πρέσπα φάνηκε να επηρεάζεται εντονότερα, σε σχέση με τη βαθιά λίμνη Μεγάλη Πρέσπα, από το φαινόμενο της κλιματικής αλλαγής και την εποχικότητα.
**Table of Contents**

1. Introduction to water quality modelling .................................................. 29
   1.1. Overview of hydrological and water quality models .................................. 29
   1.2. The role of Artificial Neural Networks (ANNs) in Eco-hydrological Modelling with an emphasis on algal production .......................................................... 32
   1.3. References ................................................................................................. 37

2. Feedforward Neural Networks .................................................................. 44
   2.1. An overview of Artificial Neural Networks as a method of Artificial Intelligence ................................................................. 44
   2.2. Feed-forward multi-layer perceptron neural network .................................. 46
   2.3. ANN topology selection ............................................................................. 53
   2.4. Measured data set pretreatment .................................................................. 57
   2.5. Performance Evaluation of the ANNs ......................................................... 63
   2.6. Sensitivity Analysis .................................................................................... 67
   2.7. Two-way sensitivity analysis ..................................................................... 69
   2.8. References ................................................................................................. 70

3. Unsupervised Neural Networks .................................................................. 81
   3.1. Self-Organizing Map (SOM) neural networks ............................................. 81
   3.2. The structure and basic steps of the SOM algorithm ..................................... 84
   3.3. U-matrix and Component Planes ................................................................. 97
   3.4. The Cluster Structuring Index .................................................................... 103
   3.5. Clustering of the SOM .............................................................................. 107
   3.6. Application of the SOM neural network for prediction .............................. 113
   3.7. References ................................................................................................. 116

4. Artificial Neural Networks and other popular modelling techniques. ...... 125
   4.1. Applications of Statistical Techniques in water quality modelling .............. 125
   4.2. Correlation Coefficient ............................................................................ 126
   4.3. Principal Components Analysis ................................................................ 128
   4.4. Factor Analysis ......................................................................................... 130
   4.5. Cluster analysis ....................................................................................... 133
   4.6. Multiple Linear Regression ....................................................................... 141
   4.7. Studies combining ANN models with other modelling methodologies ........ 142
   4.8 ANN models vs other modelling methodologies ....................................... 147
# 5. Application 1: Assessing the Contribution of the Environmental Parameters to Eutrophication with the Use of the “PaD” and “PaD2” Methods in a Hypereutrophic Lake

5. 1. Introduction .................................................................................................................. 160

5. 2. Materials and Methods ............................................................................................... 162
   5.2.1. Study Area and Data Collection .............................................................................. 162
   5.2.2. ANNs Methodologies ............................................................................................. 164
   5.2.3. ANN Model Development ....................................................................................... 168

5. 3. Results and Discussion ............................................................................................... 170
   5.3.1. ANN’s Simulation Results ....................................................................................... 170
   5.3.2. Implications for Management and Restoration ...................................................... 175

5. 4. Conclusions .................................................................................................................. 178

5. 5. References ..................................................................................................................... 179

# 6. Application 2: Evaluating the contributing environmental parameters associated with eutrophication in a shallow lake by applying Artificial Neural Networks Techniques

6. 1. Introduction .................................................................................................................. 186

6. 2. Materials and Methods ............................................................................................... 188
   6.2.1. Monitoring Area and Monitored Data. ................................................................... 188
   6.2.2. Artificial Neural Networks Methodologies. ............................................................ 189
   6.2.3. ANN Model Development. ....................................................................................... 192

6. 3. Results and Discussion ............................................................................................... 194

6. 4. Conclusions .................................................................................................................. 197

6. 5. References ..................................................................................................................... 199

# 7. Application 3: Assessment of the Eutrophication-Related Environmental Parameters in Two Mediterranean Lakes by Integrating Statistical Techniques and Self-Organizing Maps

7. 1. Introduction .................................................................................................................. 204

7. 2. Materials and Methods ............................................................................................... 207
   7.2.1. Study Area and Data Collection .............................................................................. 207
   7.2.2. Statistical Methods and Theoretical Background ................................................... 210
   7.2.3. Self Organizing Map Theory .................................................................................... 211

7. 3. Results .......................................................................................................................... 214
7.3.1. PCA and Cluster Analysis Results ................................................................. 214
7.3.2. SOM Algorithm Results .............................................................................. 217
4. Discussion ........................................................................................................ 221
7.5. Conclusions ................................................................................................... 227
7.6. References ...................................................................................................... 228
1. Introduction to water quality modelling.

1.1. Overview of hydrological and water quality models.

Hydrological and water quality models can be considered as a mathematical representation of an aquatic ecosystem, that is trying to describe/mimic a dynamical or static phenomenon of the aquatic system’s by integrating several related physical, biological and chemical variables. An example of a mathematical model applied in a limnological application is described in a study of Mellios et al. (2015), where several abiotic and biotic parameters are acting as input variables for the model and the model’s structure is analytically represented in a diagram (see Figure 1.1). The hydrological and water quality models are mainly used to assess the impacts of climate change, land use, land and crop management practices on the quality and quantity of water resources (Marias et al., 2012).

The significance of water quality models is huge and multifold, since they can contribute into reducing the cost of expensive experiments by predicting an expensive to measure parameter or by simulating events taking place into inaccessible sites, because of serious environmental pollution issues (Wang et al., 2013). Furthermore, the water quality models can predict the interaction among the water quality parameters and the impact/effect of each environmental parameter on the simulated output parameter (Hadjisolomou et al., 2016), providing the associated authorities with possible managing scenarios for handling a water quality problem.
Figure 1.1. Structure of the PCLake model, that is employing a set of coupled differential equations, applied in Lake Karla (Greece) where several biotic and abiotic parameters are included. The interactions among biotic and abiotic parameters are shown with the use of arrows (figure taken by Melliós et al. 2015).

A well constructed water quality model must meet some basic criteria in order to be considered suitable for modelling a particular purpose, such as those described below (Hajigholizadeh et al., 2018):

- Take into consideration the dataset’s nature requirements, that are associated with the phenomenon to be simulated.
- The fundamental assumptions that are related with the model’s construction.
• The performance of the model based on accuracy and validity terms.
• The components (or variables) which are the constituting elements of the model.
• The capabilities given by the model, like examining variables interactions or contributing significance based on terms such as sensitivity analysis.
• The scales that the modeled outputs can be examined and in order to be reliable/meaningful.
• Hardware requirements for simulating the model, considering model complexity and available data volume.

The need for scientifically based methods for managing water resources, led to the creation of many water quality models (Tsakiris and Alexakis, 2012). A review of water quality models is presented in the study of Gao and Li (2014), where among them the three most well-known models in hydrological studies are mentioned to be: the Soil Water Assessment tool (SWAT); the Water Quality Analysis Simulation Program (WASP); the MIKE 11. A literature-based presentation of the SWAT model and its applications can be found in the studies of Gassman et al. (2007); Bressiani et al. (2015); Stefanidis et al. (2016). The WASP model and its applications are presented in the studies of Ernst and Owens (2009); Tang et al. (2013); Wang et al. (2014). The MIKE 11 model is used in the studies of Panda et al. (2010); Pramanik et al. (2010); Doulgeris et al. (2015).

Beside the above-mentioned models or the several others well known hydrological models like the QUALs; HSPF; CE-QUAL-W2; ELCOM-CAEDYM; EFDC (Wang et al., 2013; Gao and Li, 2014), new tendencies for future hydrological modelling applications have been raised and are analyzed as following (Gao and Li, 2014):
I. Combination models: they are consisted of a combination of two or more individual models that are simulating a different phenomenon.

II. Artificial Intelligence based models: these non-mechanistic models are models like artificial neural networks; genetic algorithms; support vector machines and fuzzy reasoning.

III. System integration models: the integration of remote sensing system (RSS); geographical information system (GIS); global position system (GPS), which are called the 3S and can collect and process massive amounts of hydrological data.

1.2. The role of Artificial Neural Networks (ANNs) in Eco-hydrological Modelling with an emphasis on algal production.

Artificial Neural Networks (ANNs) are widely used in hydrological modelling during the last decades, mainly because their theoretical background is evolving and the computational processing power is increasing (Karul et al., 2000). Initially only few applications of the ANNs dealing with environmental and ecological aspects were reported at the beginnings of the 90s (Lek and Guegan, 1999). The application of ANNs for simulating trophic function and harmful algal blooms is also a relative recent task (Teles et al., 2006). Some of the first applications that, simulated algal production with the use of an ANN model, are the studies of French and Recknagel, (1994); Maier and Dandy (1996); Scardi, (1996); Recknagel et al., (1997); Maier and Dandy (1997); Yabunaka et al., (1997); Maier et al., (1998); Karul et al., (1999); Scardi and Harding, (1999).

Most of the water quality studies, that are using ANNs as modelling tools, are dealing with limnological or riverine applications, while ANN applications examining coastal systems are very scarce (Chau, 2006). Specifically, for eutrophication related problems and their
catastrophic effects, the use of ANN’s predictions can prevent or minimize the effects of any possible HABs (Devillers, 2009). The ANN’s well known abilities to model complex and non-linear relationships is ideal for eutrophication modelling. Based on these abilities ANNs are considered as a useful management tool regarding lake eutrophication (Karul et al., 1999).

Besides the predicting ability of ANNs to simulate algal productivity with a good accuracy, ANNs are good at examining the effect of the related water quality parameters based on their ability to associate them with the function of the algal biomass (Kilic et al., 2006). This is can be archived based on several mathematical relationships that are examining the sensitivity analysis for each environmental parameter and the associated impact on the production of algal biomass (Hadjisolomou et al., 2017). Generally, the results of the sensitivity analysis of a trained ANN model are providing the modelers with useful information about the water quality parameters interrelationships in an ecosystem (Recknagel, 2001). A more detailed presentation of the sensitivity analysis methods is given in a following chapter. As stated by Chen and Zhang (2009) the ANNs can be meaningfully applied for the examination of effect-relations and the impact of several environmental parameters on a complex environmental system, in contrast with other conventional methods.

Besides the usage of sensitivity analysis, in their study Millie et al. (2012) are applying ANNs to model the microalgal abundance and are proposing a more sophisticated way to exam the relationships between the associated water quality parameters and the Chl-a. As stated by the above authors, this proposed method is based on a pedagogical knowledge extraction method by using the multi-dimensional response surfaces and by that way it was possible the creation of mathematical equations of iterative predictive outcomes of Chl-a. In another ANN study of Millie et al. (2014), a similar methodology is used to produce mathematical equations between the Chl-a and the associated water quality parameters and to visualize the Microcystins along with the associated environmental parameters interactions.
These two mentioned studies are presented in a more detailed way in a following chapter, since they provide an alternative mathematical way to associate the Chl-\(a\) and the several water quality parameters.

Comparing classical modelling techniques used in hydrological modelling (including water quality related problems), it is concluded that the ANN modelling techniques are superior and produce more reliable results. This is attributed to the fact that the data driven modelling techniques, like ANNs, are more advantageous for water quality modelling comparing with the classical process-based modelling techniques. Since the process-based techniques are usually requiring a large amount of monitored data samples and a large number of environmental variables (Karul et al., 2000), while the data driven modelling techniques are requiring fewer input parameters than process-based techniques and are computationally faster (Palani et al., 2008). Generally the process-based modelling techniques are less reliable than the data driven techniques, because they require too many data and parameters for calibration and their results are usually restricted to the time and sites that they were developed within (Chen et al., 2010). Based on the relative literature examination it is found that the ANNs are better predictors regarding water quality modelling comparing with general linear models, however some exceptions exist when a general linear regression model outperforms an ANN model (Ozesmi et al., 2006).

In the study of Jorgensen (2008) an overview of the major model types that are used in ecological modelling is presented: bio-chemical and bio-energetics dynamic models; static models; population dynamic models; structurally dynamic models; fuzzy models; artificial neural network; individual-based models and cellular automata; spatial models; ecotoxicological models; stochastic models; hybrid models. A comparison of the ANNs against the rest modelling methods is presented by the author and the main advantages of the ANNs are given as follow (Jorgensen, 2008):
• ANNs are easy to be applied.
• They are not restricted of the data set heterogeneity.
• Can be used in cases when other modelling methods fail to be applied.
• Based on the test set usage they are evaluated and give a reliable indication of the certainty.

Where their main disadvantages of the ANNs comparing with the rest modelling techniques are associated with the following as stated by Jorgensen (2008):

• Bio-geo-chemical models that are based on the conservation principles cannot be replace by the ANNs.
• ANNs are not able to examine the causality of the simulated processes, unless they are combined or introduced with an algorithm.
• ANNs sometimes are having limited accuracy.

Even though ANNs are having better modelling results than other modelling methods, sometimes even the ANNs fail to produce good enough output results, however ANNs are flexible modelling techniques and an experienced modeler might produce improved outputs. For example this problem is examined in the study of Scardi (2001), which is dealing with the problem of limited available data when modelling with ANN techniques the phytoplankton primary production and for this purpose co-predictors (defined as variables that are correlated with the ones to be predicted, but not directly related to them) and the metamodelling procedure (when using additional information derived from other models in order to train, validate and test an ANN model) are used in order to improve previously constructed ANNs.
and as it was concluded by the author the ANNs are considered black-box models that their performance has the potential to be improved after applying some “modelling tricks”.

ANN models can be successfully used for time series prediction of a phenomenon. For the purpose of time series modelling of the eutrophication problem in lakes, time lagged input environmental parameters are used as an ANN model’s inputs. By that way the constructed ANN model can act as a warning tool against a possible HAB event, when Chl-α levels are predicted to reach high values. Some representative studies of ANN’s time series modelling of eutrophication can be found in the studies of Belgrano et al. (2001); Lee et al. (2003); Li et al. (2007); Hadjisalomou et al. (2017).

However, the majority of ANNs that are used in hydrological/ecological modelling are not using time lagged parameters and their main objective is to simulate (mimic) a hydrological event based on a number of associated environmental parameters. Several applications exist on the eco-hydrological field, something that is closely related to the fact that the ANNs are having higher accuracy in the prediction of algal concentration than other modelling approaches (Li et al., 2007). Several studies that have compared the simulated results of an ANN model with those of a linear regression model, have verified that the ANNs are better predictors. Some examples of such studies are the ones of Lek et al., (1996); Karul et al., (2000); Ozesmi et al., (2006); Mas and Ahlfeld, (2007); Adamowski and Karapataki, (2010); Abba et al., (2017)

Another category of ANNs broadly used in the eco-hydrological field is the one of the unsupervised ANNs, known as the Kohonen Self-Organizing Maps (SOMs). These SOMs are mainly used for clustering/patterning environmental data and for exploratory data analysis (data mining) of the environmental data set. As stated by Park et al. (2003) usually the conventional multivariate analyses have been used, but ANNs are more suitable for ecological
patterning because the interactions in the data set, which is consisted of many species and sampling areas, are nonlinear and complex.

An example of a SOM applied in eco-hydrological modelling is found in the study of Lu and Lo (2002), where a trophic state classifier was constructed based on a SOM model, aiming to diagnose the water quality of the Fei-Tsui Reservoir (Taiwan) during the monitoring period 1987-1995 and compared the simulated SOM results with those of the Carlson Index. In another application given by Choi et al. (2014) the SOM model was used and hydrochemical data (e.g. pH, Cd, Hg, Ca, Na, K, NO3) of bedrock groundwater of 299 sampling stations in South Korea were clustered into two main groups and by that way the water quality and the anthropogenic contamination was evaluated. In the study of Hadjisolomou et al. (2018) the SOM model is applied into the transboundary Prespa Lakes (examining only the Greek part) aiming to explore the interrelationships among several water quality parameters (like TP; DIN, WT; Chl-a) that are associated with algal production. By that way the constructed SOM model acted as an eutrophication control management tool and also helped to understand the underlying mechanisms associated with trophic productivity.

1.3. References

Abba, S.I.; Hadi, S.J.; Abdullahi, J. River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. Procedia Computer Science 2017, 120, 75-82.

Adamowski, J; Karapataki, C. Comparison of multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. Journal of Hydrologic Engineering 2010, 15, 729-743.
Belgrano, A.; Malmgren, B.A.; Lindahl, O. Application of artificial neural networks (ANNs) to primary production time-series data. *Journal of Phytoplankton Research* **2001**, *23*, 651-658.

Bressiani, D. A.; Gassman, P. W.; Fernandes, J. G.; Garbossa, L. H. P.; Srinivasan, R.; Bonumá, N. B.; et al. Review of Soil and Water Assessment Tool (SWAT) applications in Brazil: Challenges and prospects. *Int J Agric & Biol Eng* **2015**, *8*, 9–35.

Chau, K.-W. A review on integration of artificial intelligence into water quality modelling. *Marine Pollution Bulletin* **2006**, *52*, 726-733.

Chen, D.; Lu, J.; Shen, Y. Artificial neural network modelling of concentrations of nitrogen, phosphorus, and dissolved oxygen in a non-point source polluted river in Zhejiang Province, southeast China. *Hydrological Processes* **2010**, *24*, 290-299.

Chen, L.-H.; Zhang, X.-Y. Application of Artificial Neural Networks to classify Water quality of the Yellow River. In *Fuzzy Information and Engineering*; Cao, B.Y.; Zhang, C.Y.; Li, T.F. Eds; Springer -Verlag: Berlin, Germany, 2009.

Choi, B.-Y.; Yun, S.-T.; Kim, K.-H.; Kim, J.-W.; Kim, H.M et al. Hydrochemical interpretation of South Korea groundwater monitoring data using Self-Organizing Maps. *Journal of Geochemical Exploration* **2014**, *137*, 73-84.

Devillers, J. Artificial Neural Network Modeling of the Environmental Fate and Ecotoxicity of Chemicals. In *Ecotoxicology Modeling*; Devillers, J. Eds.; Springer -Verlag: Boston, MA, USA, 2009.

Doulgeris, C.; Georgiou, P.; Papadimos, D.; Papamichail, D. Water allocation under deficit irrigation using MIKE BASIN model for the mitigation of climate change. *Irrig Sci* **2015**, *33*, 469-482.
Ernst, M.R.; Owens, J. Development and application of a WASP model on a large Texas reservoir to assess eutrophication control. *Lake and Reservoir Management* **2009**, *25*, 136-148.

French, M; Recknagel, F. Modelling of algal blooms in freshwaters using artificial neural networks. In *Computer techniques in environmental studies*; Zannetti, P. Eds.; WIT Press, Boston, USA, 1994; Volume 6, pp 87-94.

Gao, L.; Li, D. A review of hydrological/water -quality models. *Front. Agr. Sci. Eng.* **2014**, *1*(4), 267-276.

Gassman, P.W.; Reyes, M.R.; Green, C.H.; Arnold, J.G. THE SOIL AND WATER ASSESSMENT TOOL: HISTORICAL DEVELOPMENT, APPLICATIONS AND FUTURE RESEARCH DIRECTIONS. *Transactions of the ASABE* **2007**, *50*, 1211-1250.

Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G. Papastergiadou, E. Assessing the Contribution of the Environmental Parameters to Eutrophication with the Use of the “PaD” and “PaD2” Methods in a Hypereutrophic Lake. *International Journal of Environmental Research and Public Health* **2016**, *13*(8):764, doi: 10.3390/ijerph13080764

Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G. Papastergiadou, E. Evaluating the contributing environmental parameters associated with eutrophication in a shallow lake by applying artificial neural networks techniques. *Fresenius Environmental Bulletin* **2017**, *26*(5):3200-3208

Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G. Papastergiadou, E. Assessment of the Eutrophication-Related Environmental Parameters in Two Mediterranean Lakes by Integrating Statistical Techniques and Self-Organizing Maps. *International Journal of*
Hajigholizadeh, M; Melesse, AM; Fuentes, HR; Erosion and Sediment Transport Modelling in Shallow Waters: A Review on Approaches, Models and Applications *International Journal of Environmental Research and Public Health* **2018**, 15(3), doi: 10.3390/ijerph15030547.

Jorgensen, S.E. Overview of the model types available for development of ecological models. *Ecological Modelling* **2008**, 215, 3-9.

Karul, C.; Soyupak, S.; Yurteri, C. Neural networks as a management tool in lakes. *Hydrobiologia* **1999**, 408/409, 139-144.

Karul, C.; Soyupak, S.; Cilesiz, A.F.; Akbay, N.; Germen, E. Case studies in the use of neural networks in eutrophication modeling. *Ecological Modelling* **2000**, 134, 142-152.

Kilic, H.; Soyupak, S.; Gurbuz, H.; Kivrak, E. Automata networks as preprocessing technique of artificial neural network in estimating primary production and dominating phytoplankton levels in a reservoir: An experimental work. *Ecological Informatics* **2006**, 1, 431-439.

Lee, J.; Huang, Y.; Dickman, M.; Jayawardena, A.W. Neural network modelling of coastal algal blooms. *Ecological Modelling* **2003**, 159, 179-201.

Lek, S.; Delacoste, M.; Baran, P.; Dimopoulos, I.; Lauga, J.; Aulagnier, S. Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling* **1996**, 90, 39-52.

Lek, S.; Guegan, J.F. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* **1999**, 120, 65-73.
Li, H.; Hou, G.; Dakui, F.; Xiao, B.; Song, L. et al. Prediction and elucidation of the population dynamics of *Microcystis* spp. in Lake Dianchi (China) by means of artificial neural networks. *Ecological Informatics* 2007, 2, 184-192.

Lu, R.-S.; Lo, S.-L. Diagnosing reservoir water quality using self-organizing maps and fuzzy theory. *Water Research* 2002, 36, 2265-2274.

Maier, H.R.; Dandy, G.C. The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters. *Water Resources Research* 1996, 32, 1013-1022. Maier, H.R.; Dandy, G.C. Modelling cyanobacteria (blue-green algae) in the River Murray using artificial neural networks. *Mathematics and Computers in Simulation* 1997, 43, 377-386.

Maier, H.R.; Dandy, G.C.; Burch, M.D. Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. *Ecological Modelling* 1998, 105, 257-272.

Mas, D.M.L.; Ahlfeld, D. Comparing artificial neural networks and regression models for predicting faecal coliform concentrations. *Hydrological Sciences Journal* 2007, 52, 713-731.

Mellios, N.; Kofinas, D.; Laspidou, C.; Papadimitriou, T. Mathematical Modelling of Trophic State and Nutrient Flows of Lake Karla using the PCLake Model. *Environ. Process.* 2015, 2, 85-100.

Millie, D.F.; Weckman, G.R.; Young II, W.A.; Ivey, J.E.; Carrick, H.J.; Fahnenstiel, G.R. Modeling microalgal abundance with artificial neural networks: Demonstration of a heuristic ‘Grey-Box’ to deconvolve and quantify environmental influences. *Environmental Modelling & Software* 2012, 38, 27-19.

Millie, D.F.; Weckman, G.R.; Fahnenstiel, G.R.; Carrick, H.J.; Ardjmand, E. et al. Using artificial intelligence for CyanoHAB niche modeling: discovery and visualization of
Microcystis–environmental associations within western Lake Erie. Canadian Journal of Fisheries and Aquatic Sciences 2014, 71, 1642-1654.

Moriai, D.N.; Wilson, B.N.; Douglas-Mankin, K.R.; Arnold, J.G.; Gowda, P.H. HYDROLOGIC AND WATER QUALITY MODELS: USE, CALIBRATION, AND VALIDATION. Transactions of the ASABE 2012, 55, 1241-1247.

Ozesmi, S.L.; Tan, C.O.; Ozesmi, U. Methodological issues in building, training, and testing artificial neural networks in ecological applications. Ecological Modelling 2006, 195, 83-93.

Palani, S.; Liong, S.-Y.; Tkalich, P. An ANN application for water quality forecasting. Marine Pollution Bulletin 2008, 56, 1586-1597.

Panda, R.K.; Pramanik, N.; Bala, B. Simulation of river stage using artificial neural network and MIKE 11 hydrodynamic model. Computers & Geosciences 2010, 36, 735-745.

Park, Y.-S.; Verdonschot, P.F.M.; Chon, T.-S.; Lek, S. Pattern and predicting aquatic macroinvertebrate diversities using artificial neural network. Water Research 2003, 37, 1749-1758.

Pramanik, N.; Panda, R.K.; Sen, D. One Dimensional Hydrodynamic Modeling of River Flow Using DEM Extracted River Cross-sections. Water Resources Management 2010, 24, 835-852.

Recknagel, F.; French, M.; Harkonen, P.; Yabunaka, K.-I. Artificial neural network approach for modelling and prediction of algal blooms. Ecological Modelling 1997, 96, 11-28.

Recknagel, F. Applications of machine learning to ecological modelling. Ecological Modelling 2001, 146, 303-310.

Scardi, M. Advances in the neural network modeling of phytoplankton primary production. Ecological Modelling 2001, 146, 33-45.
Scardi, M. Artificial neural networks as empirical models for estimating phytoplankton production. *Marine Ecology Progress Series* 1996, *139*, 289-299.

Scardi, M.; Harding Jr, H.W. Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological Modelling* 1999, *120*, 213-223.

Stefanidis, K.; Panagopoulos, Y.; Psomas, A.; Mimikou, M. Assessment of the natural flow regime in a Mediterranean river impacted from irrigated agriculture. *Science of the Total Environment* 2016, *573*, 1492-1502.

Tang, P.-K.; Huang, Y.-C.; Kuo, W.-C.; Chen, S.-J. Variations of model performance between QUAL2K and WASP on a river with high ammonia and organic matters. *Desalination and Water Treatment* 2013, *52*, 1193-1201.

Teles, L.O.; Vasconcelos, V.; Pereira, E.; Saker, M.L. Time Series Forecasting of Cyanobacteria Blooms in the Crestuma Reservoir (Douro River, Portugal) Using Artificial Neural Networks. *Environmental Management* 2006, *38*, 227-237.

Tsakiris, G.; Alexakis, D. Water quality models: An overview. *European Water* 2012, *37*, 33-46.

Wang, F.; Li, Y.; Yang, L.; Sun, Z. Application of WASP model and Gini coefficient in total mass control of water pollutants: a case study in Xicheng Canal, China. *Desalination and Water Treatment* 2014, *57*, 2903-2916.

Wang; Q.; Li, S.; Jia, P.; Qi, C.; Ding, F. A review of surface water quality models. *The Scientific World Journal* 2013, doi.org/10.1155/2013/231768

Yabunaka, K.-I.; Hosomi, M.; Murakami, A. A novel application of a backpropagation artificial neural network model formulated to predict algal bloom. *Water Science and Technology* 1997, *36*, 89-97.
2. Feedforward Neural Networks

2.1. An overview of Artificial Neural Networks as a method of Artificial Intelligence

Different Artificial Intelligence (AI) techniques are suitable for eco-hydrological modelling, since the interactions of the hydrological parameters are complex and non-linear (Ay and Kisi, 2014). Several modelling studies exist that are applying AI methods in order to model water resources e.g. Chaves et al. (2004); Shang (2005); Chen et al (2008); Wang et al. (2009); Turan and Yurdusev (2009); Emamgholizadeh et al. (2014); Seo et al. (2015); Zaji and Bonakdari (2018); Ceccaroni et al. (2018). In the study of Chau (2006) an extensive review of several AI techniques like ANNs, genetic algorithms, knowledge-based systems and their integration into eco-hydrological modelling are presented. The most popular AI techniques that are found in eco-hydrological modelling studies are the ANNs and fuzzy controllers (Kruszewski et al., 2008; Han and Qiao, 2011). In the study of Millie et al. (2012) it is stated that “ANNs are a core form of Artificial Intelligence models that discern complex associations among variables through iterative and repetitive data presentation”.

Artificial Neural Networks are considered to be a very promising machine learning method, since they can simulate non-linear relationships and to assess interactions between the environmental variables (Lek and Guegan, 1999; DeWeber and Wagner, 2014), while in many cases these environmental parameters are interrelated (Ahmed, 2014). Furthermore, the fact that the application of an ANN does not require a priori knowledge of the underlying mechanism/process, established the ANNs to be a suitable tool for handling various hydrological problems. Therefore the number of ANN’s applications in eco-hydrological studies is increasing rapidly over the last decades (Faruk, 2010; Basant et al., 2010; Sarkar and
The fact that the ANNs can obtain high simulation and forecasting accuracy in eco-hydrological modelling established them as a management tool (An et al., 2015).

Generally, the ANNs are considered as ‘black box’ models by many researchers (Lek and Guegan, 1999; Kalteh et al., 2008; Ahmed, 2014), requiring no detailed information about the system (Ahmed, 2014). However, the term ‘Grey-Box’ is most suitable according to Millie et al. (2012), because based on heuristic knowledge extraction technique from a trained ANN the environmental interactions for the Chl-\textit{a} parameter can be quantified through the summation of the related response-surface equations.

A very simple/primary definition of the ANN is that an ANN can be considered as a computing system, which is consisted of a highly interconnected set of simple information processing elements (units), analogous to a biological neuron (Sarkar and Pandey, 2015). The aim of an ANN is to create a model of the data-generated weight process, which can generalize and predict outputs from data that haven’t been presented before (unknown / unseen) to the ANN model (Lee et al., 2013). ANNs are inspired by the ability of human beings to perform well complex tasks through the processing of the biological neural system (Han et al., 2012). As in nature, the biological neural system function is determined by connections through the elements/neurons (Sener et al., 2012).

The importance of neurons is significant for ANN modelling and as is stated by Munakata, (2008): “the way a natural neuron is the building block of the brain, the basic element of every ANN is an artificial neuron”. A figure of a biological neuron is provided below (see Figure 2.1), where the comparison/parallelization with an artificial neuron is observed. The neuron is consisted by three elements: the dendrites, cell body and the axon. The parallelization between a biological neuron and an artificial neuron is that the connections between the nodes can be parallelized with the axon and dendrites; the biological synapses can
be parallelized with the connection weights; the threshold of the artificial neuron is approximating the activity in the cell body/soma (Sanchez Mesa et al., 2005).

Figure 2.1. Parallelization between a biological neuron and an artificial neuron (taken by Sanchez Mesa et al., 2005).

2.2. Feed-forward multi–layer perceptron neural network

ANNs can be categorized into two groups, supervised and unsupervised. In the case of ANNs with supervised learning the given data is consisted of inputs and associated output–targets. In contrast, ANNs with unsupervised learning, in their given data set have no output–targets, therefore their aim is to discover patterns among inputs (Engelbrecht, 2007). The Kohonen’s Self-Organizing Map (SOM) is the most famous unsupervised ANN, as it is analyzed by Kohonen, (1990). A detailed presentation of SOM neural networks is given in Chapter 3.
The ANNs have been described as parallel-distributed information processing systems that are mimicking the way the biological neural networks are processing the information and are having the ability to process/solve large-scale complex problems (Kalteh et al., 2008). Because the ANNs are highly parallel architectures (meaning that their numerous independent operations can be executed simultaneously), the ANNs are the most preferred technique today for high speed processing of huge data sets (Sarkar and Pandey, 2015). The learning mechanism of ANNs was defined by Ball and De la Rosa (2006) as: “a computational mechanism able to acquire, represent, and compute a weighting (or mapping) from one multivariate space of information to another, given a set of data representing that mapping”.

ANNs with skip layer connections are ANNs with direct connections between the input layer neurons and the output layer neurons, therefore linear relationships between the input variables and the simulated output are resulting (DeWeber and Wagner, 2014). Such ANN models, which have no hidden layers, are considered analogous to a linear model, while the increasing nonlinearities in the modeled relationships are analogous to the increasing number of hidden neurons (Cheng and Titterington, 1994; DeWeber and Wagner, 2014).

The most commonly used ANN in eco-hydrological modelling is the feed-forward multi-layer Perceptron (MLP) with back-propagation training algorithm (Maier and Dandy, 2000; Kalteh et al., 2008, Najah et al., 2011). The MLP neural networks are closely related to statistical models and are the most suitable type of ANNs for forecasting a phenomenon (Rumelhart et al., 1986; Najah et al., 2011). Additionally, as it is stated by Hsu et al. (1995) the three-layer feed-forward MLP can simulate successfully any of the real-world functional relationships, which may have a poorly defined form and also may have an unknown complexity. This means that the feed-forward neural network with one layer can approximate any smooth, measurable function by associating the output variable with the input variables with the use of connecting weights and transfer functions (Basant et al., 2010; Areerachakul et
Therefore, the multilayer feed-forward neural networks can be characterized as “universal approximators” capable of learning any continuous function (Hornik et al., 1989; Cybenko, 1989).

Feedforward networks are suitable for function approximation (Gupta et al., 2003). A feedforward ANN, as it can be derived by its characterization, has an acyclic topology with no feedbacks loops among its layers and it is used for approximating non-linear mappings among inputs and outputs (Hu and Hwang, 2002). An ANN is divided by at least three layers. The input layer that imports the input parameters to the network, then one or more hidden layers and the output layer that gives the result. Each layer is consisted of neurons, also called nodes, as a neural network resembles a graph (Kasabov, 1998). The architecture of a feedforward neural network is graphically illustrated as in the figure below (Figure 2.2).

For every neuron there is synaptic weight that connects the specific neuron with every neuron of the next layer. Each neuron has a number of inputs, with exception those of the first layer, and one output (Dedecker et al., 2004). Feedforward networks are divided into at least three layers and each layer is consisted by neurons. The first layer is the input layer, at least one intermediate hidden layer and the output layer that produces the result. Each neuron of a layer is connected with all the neurons of the next layer with a synaptic weight. Aggregation is performed for every neuron on its weighted inputs from the previous layer and an output is yielded through a transfer function or else activation function (Veeenturf, 1995; Recknagel et al, 1997; Gupta et al., 2003). The output value of the jth neuron \( o_j \) is given by the equations as described by Dedecker et al. (2004):

\[
o_j = f(u_j) \quad \text{(Eq. 2.1)}
\]

and

\[
u_j = \sum w_{ij} x_i + z_j \quad \text{(Eq. 2.2)}
\]
**Figure 2.2.** Schematic representation of a feedforward ANN with a hidden layer with $j$ neurons on it, $n$ inputs and one output. The blue arrows are describing the way the sum of weighted inputs from the input layer with the use of the activation function becomes an output ($o_j$) for the $j$th-neuron. Where the $w_{ij}$ are the synaptic weights and $x_1, x_2, \ldots, x_n$ are the input variables.

where $f$ is the transfer function, $x_i$ is the input from $i$th neuron belonging to the immediate previous layer, $w_{ij}$ is the synaptic weight that connects $x_i$ with the $j$th neuron and $z_j$ a bias term. The output of each neuron is computed and propagated through the next layer until the last layer, producing a network output that is compared with the given output (Ghalkhani et al. 2013).
ANNs are trained with the use of a learning algorithm. Feedforward networks can be trained with the use of back-propagation algorithm’s variations. During the learning procedure the synaptic weights are adjusted so they are minimizing an error function, usually the mean square difference between the predicted and the given output (Vilas et al., 2011). The ANN’s training process is repeated until converge and the ANN is able to produce good outputs when is given unseen (unknown) input data (Kalteh et al., 2008). After the ending of the training phase, the ANN should be able to simulate the input data and produce reliable output values (Lee et al., 2013).

Back-propagation is the most famous and frequently used algorithm for training feed-forward MLP (Haykin, 1999; Kalteh et al., 2008; Areerachakul et al., 2013; Lee et al., 2013; Zhang and Ding, 2017). The Levenberg-Marquardt algorithm, which is a modification of the back-propagation algorithm, is preferred for the training of moderate-sized feed-forward ANNs, because it is reported to be the fastest method (Demuth and Bealy, 2002; Simon, 2005; Cho et al., 2014). According to Ding et al. (2014) the application of the Levenberg-Marquardt optimization algorithm may shorten the learning time and improve the learning speed, because ANNs trained with the back-propagation algorithm have problems related with the existence of local minimum in convergence learning and slow convergence rates. Some variations of the backpropagation algorithm are the Levenberg–Marquardt algorithm, back-propagation with momentum, BFGS Quasi-Newton algorithm, resilient backpropagation algorithm and scaled conjugate gradient algorithm (Hagan et al., 1996; Demuth and Beale, 2002).

The Levenberg-Marquardt algorithm is a very popular optimization algorithm used for training the neural networks and widely used in eco-hydrological ANN modelling (e.g. Najah et al., 2011; Millie et al., 2012; Wu et al., 2014; Hadjisalomou et al., 2016). The Levenberg-Marquardt algorithm converges faster among all backpropagation algorithm modifications, for training moderate-sized feedforward neural networks, up to several hundred weights (Demuth
and Beale, 2002), something that is providing Levenberg-Marquardt algorithm against other ANN's training algorithms a clear benefit when simulating large databases. The mathematical relationship describing the Levenberg-Marquardt algorithm is given as below:

\[ x_{k+1} = x_k - [J^TJ + \mu I]^{-1}J^Te \]  \hspace{0.5cm} (Eq. 2.3)

where \( J \) stands for the Jacobian matrix of the first derivatives of the network's errors with respect to the weights and biases and \( e \) is a vector of network errors. \( I \) is the identity matrix and \( J^T \) is the transpose matrix of matrix \( J \). The \( \mu \) is a scalar, that when approximates zero, then the Levenberg-Marquardt algorithm becomes the Newton's algorithm and when \( \mu \) is large enough the Levenberg-Marquardt algorithm becomes the gradient descent method (Hagan et al., 1996; Demuth and Beale, 2002).

The most popular activation functions used for ANNs applications are the linear activation function and the sigmoid (which is nonlinear) activation function (Dawson and Wilby, 2001; Rankovic et al., 2010; Ay and Kisi, 2014). The non-linear activation functions are giving the ANN the ability to simulate well the non-linear relationships between the input variables and the simulated/predicted output value (Kmet and Kmetova, 2015). The most popular activation functions are having the following mathematical forms (Demuth and Bealy, 2002; Rankovic et al., 2010):

- Linear activation function with the following mathematical form:

\[ f(n) = n \]  \hspace{0.5cm} (Eq. 2.4)

The linear activation function can be graphically represented based on Figure n, where its linear nature can be observed.
Figure 2.3. Graphical representation of the linear activation function (partially modified from Demuth and Bealy, 2002).

- Logistic activation function, which is a sigmoid function and is having the following mathematical form:

\[
f(n) = \frac{1}{1+e^{-n}} \quad \text{(Eq. 2.5)}
\]

The logistic activation function can be graphically represented based on Figure n, where its nonlinear nature can be observed.

Figure 2.4. Graphical representation of the logistic activation function (partially modified from Demuth and Bealy, 2002).
- Hyperbolic tangent activation function, which is a sigmoid function and is having the following mathematical form:

\[ f(n) = \frac{1-e^{-n}}{1+e^{-n}} \quad (\text{Eq. 2.6}) \]

The hyperbolic tangent function can be graphically represented based on Figure n, where its nonlinear nature can be observed.

![Graphical representation of the hyperbolic tangent activation function](image)

**Figure 2.5.** Graphical representation of the hyperbolic tangent activation function (partially modified from Demuth and Bealy, 2002).

### 2.3. ANN’s topology selection

The number of the ANN’s hidden neurons cannot be calculated on a “magic formula” and some rules of thumb exist in order to calculate the ANN’s number of hidden neurons (Gazzaz et al., 2012). Some researchers stated that the selection of the optimal/proper number of an ANN’s hidden neurons is a trial and error procedure (e.g. Ozesmi et al., 2006; Palani et al.; 2008; Dogan et al., 2009; Basant et al., 2010; Liu and Chen, 2012), meaning that the determination of the hidden neurons number is problem depended. Meanwhile, some other researchers suggested some heuristic rules in order to find the optimal number of hidden
neurons. Some of these heuristic rules are summarized by Goethals et al. (2007) in the form of mathematical equations (Equation 2.7-2.11) as below:

- \(2 \times N_i + 1\) (Eq. 2.7)
- \(\left(\frac{2}{3}\right) \times N_i\) (Eq. 2.8)
- \(0.5 \times (N_i + N_o)\) (Eq. 2.9)
- \(0.75 \times N_i\) (Eq. 2.10)
- \(3 \times N_i\) (Eq. 2.11)

where in the above equations the symbol \(N_i\) is standing for the number of input neurons (nodes), the symbol \(N_o\) is standing for the number of output neurons.

An alternative heuristic rule for calculating the optimal neurons number is proposed by Maier et al. (1998), which is based on the combination of the following two equations (Equation 2.12-2.13):

\[N^H \leq 2N^I + 1\] (Eq. 2.12)

and

\[N^H \leq 2N^{TR} / (N^I + 1)\] (Eq. 2.13)

where \(N^H\) is the number of hidden layer neurons, \(N^I\) the number of inputs and \(N^{TR}\) the number of training samples. The maximum value of hidden neurons \((N^H)\) must be the smallest number calculated by those two equations.

Generally, it must be stated that the fewer the hidden neurons the better, since the created ANN is less prone to overfitting the unknown data and has better generalization abilities (e.g. Ozesmi et al., 2006; Palani et al.; 2008). The overfitting is a situation when the computational error of the training set is having a small value, while for the test set the ANN is having large value for the computational error (Dogan et al., 2009). This is happening because the ANN has managed to memorize well the training set, but it is unable to generalize.
well to new situations (meaning to produce good output for new-unknown data inputs). When the number of hidden neurons is too small, then the ANN model might be unable to produce good simulated outputs because the ANN doesn’t have sufficient degrees of freedom to learn the process correctly (Kisi, 2005; Basant et al., 2010; Liu and Chen, 2012). While if the number of ANN’s hidden neurons is too large, then the ANN might overfit the data and the ANN’s training may take too much time (Karunanithi et al., 1994; Kisi, 2005; Liu and Chen, 2012).

Besides finding the optimal number of neurons of an ANN, some other modelling techniques for avoiding the overfitting exist. The first one is the early stopping method and the second one is the method of regularization. In early stopping the data set is divided into three subsets: training set, validation set and test set. The network training stops when the errors for the validation set begins to rise, indicating that the network had began to overfit the data (Karul et al., 2000; Li et al., 2007). The method of regularization involves modification of the network performance function, that is the sum of squares of the network errors on the training set. The modified performance function will cause the network to have smaller weights and biases, and this will force the network response to be smoother and less likely to overfit (Demuth and Beale, 2002; Kiliç et al., 2007).

The number of the ANN’s hidden layers has also an important role for the ANN architecture. When the more hidden layers exist, then the more flexible the ANN model might become at simulating with better accuracy the data of the learning set. However, the serious risk of the ANN to be unable to simulate well unknown data is very possible, since the existence of more than one hidden layer (e.g. Figure 2.6) is increasing the number of neurons and of synaptic weights. As stated by Dogan et al. (2009) the ANNs are sensitive to the number of hidden layers.
Figure 2.6. Illustration of a multi-layer feedforward (with $L$ layers) neural network (taken by Wang et al., 2015).

Three or more hidden layers are causing unnecessary computational load to the ANN model, while an ANN architecture with one hidden layer is a more stable design. Generally, the majority of ANNs used in eco-hydrological modelling are having only one hidden layer (Basant et al., 2010), because one hidden layer is sufficient to produce arbitrarily complex outputs (Lippmann, 1987). Theoretical studies have proved that only one hidden layer is sufficient for the ANN to approximate any complex nonlinear function (Cybenko, 1989; Kisi, 2005), however the existence of more hidden layers is not prohibited (Kisi, 2005).

A graphical representation of the MSE for two different ANN architectures are presented below (see Figure 2.7). The first created ANN is having one hidden layer and the MSE is calculated for the test set and the training set. The second created ANN is having two hidden layers and the MSE is calculated for the test set and the training set. As it can be observed by the Figure 2.7, the ANN with only one hidden layer is having better performance.
(smaller SME), verifying the theory that in most cases the existence of only one hidden layer is giving better modelling results.

![Image](image-url)

**Figure 2.7.** Comparison of two ANN models’ variation of Mean Square Error (MSE) during the learning process: a) Test set’s and training set’s MSE for an ANN model with one hidden layer; b) Test set’s and training set’s MSE for an ANN with two hidden layers (partially modified from Dogan et al., 2009).

### 2.4. Measured data set pretreatment

The measured data set before been presented to the ANN must be pretreated/processed. This step is a very important since it helps to reduce the ANN model computational complexity and ensures that all the input variables will be treated by the ANN model the same way, where all these will be discussed in detail below.

The tactic to apply Principal Component Analysis (PCA) on the water quality data set before presenting the variables to the ANN model, is a widely used data-pretreatment method when using ANN models. By that way the ANN’s input parameters are selected by a larger-
initial set of candidates-parameters (Zounemat-Kermani, 2014; Zhang et al., 2015; Hadjisolomou et al., 2017). PCA is often combined with ANN modelling, because by that way dimension reduction is enabled and the model’s computational complexity is reduced, so the possibility of a model’s misconvergence and poor accuracy is eliminated (Muttil and Chau, 2007; Hadjisolomou et al., 2016). For example, in a relevant ANN modelling study of Gazzaz et al (2012) -which was using water quality variables as predictors for the water quality of the Kinta River (Malaysia)- after applying PCA to the raw dataset, the 30 initial candidates parameters served as inputs for the ANN were reduced to 23 parameters used as inputs.

After PCA is performed and decided which variables to retain as ANN inputs, then the data are pre-treated based on the following steps that are summarized by Gazzaz et al (2012) as described below:

1. The censored, missing and inconsistent measured data values were assigned an arithmetical value based on a modelling technique like linear regression, non-linear regression, ordinary least squares; or the mean value of the problematic variable for the associated monitoring period.
2. Data samples corresponding to statistical outliers and structural zeros were removed from the data set.
3. All the input variables of the ANN were standardized prior been presented to the ANN model.

The role of step 3, which is dealing with data standardization before been presented to the created ANN, is very crucial. That step is necessary in order to avoid model bias because of the different unit’s magnitude of the variables (Hadjisolomou et al., 2018). Data transformation is helping to remove any possible offsets and it ensures that the input variables
are contributing the same to the created ANN model and the distribution of the input variables are matching the distribution of the simulated output (Basant et al., 2010). The most commonly used standardization method for variables in ANN modelling is based on the following transformation, which is explained by Srinivasan et al. (1994):

\[ X_s = \left[ (b - a) \cdot \frac{X_0 - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \right] + a \quad \text{(Eq. 2.14)} \]

where in the above equation the term \( a \) is standing for the lower limit of the standardization range, the term \( b \) is standing for the upper limit of the standardization range, the term \( X_s \) is symbolizing the transformed observations of the parameter \( X \), the term \( X_0 \) is symbolizing the raw observations of the parameter \( X \), the term \( X_{\text{min}} \) is symbolizing the minimum value of the parameter \( X \), the term \( X_{\text{max}} \) is symbolizing the maximum value of the parameter \( X \). However, in most ANN modelling studies the above equation (Equation 2.14) is adjusted/modified, where the terms \( a=0 \) and \( b=1 \) and this transformation is corresponding for the well known Min-Max transformation and is having the following formula (Gazzaz et al., 2012; Antanasijevic et al., 2014):

\[ X_s = \frac{(X_0 - X_{\text{min}})}{(X_{\text{max}} - X_{\text{min}})} \quad \text{(Eq. 2.15)} \]

This is because the input variables of the ANN are standardized into the range (0,1) of the logistic sigmoid transfer/ activation function (Gazzaz et al., 2012).

It must be noted that when using the phytoplankton or primary production as input variables, then these variables must be log-transformed before been rescaled to the interval (0,1). According to Scardi and Harding Jr (1999), this is necessary because the mean square error of the ANN will be biased, if raw data are used and since the algal biomass data with
high values are containing greater sampling and measurement error, which error might dominate the ANN’s output. After the ANN’s training is over, then the simulated outputs are rescaled back to their initial (raw) form prior to post modelling computations (Basant et al., 2010; Hadjisilomou et al., 2016). Some other data transformations are used in ANN modelling. For example, in their study Tomenko et al. (2007) used a data transformation, where the variables are transformed to the range $[-1, +1]$ as follow:

$$X_s = \frac{2(X_0 - X_{min})}{(X_{max} - X_{min})} - 1 \quad \text{(Eq. 2.16)}$$

where in the above equation the term $X_s$ is symbolizing the transformed observations of the parameter $X$, the term $X_0$ is symbolizing the raw observations of the parameter $X$, the term $X_{min}$ is symbolizing the minimum value of the parameter $X$, the term $X_{max}$ is symbolizing the maximum value of the parameter $X$. Beside the mentioned above data transformations, several other used normalization techniques are used for data pre-treatment before presenting them to the ANN model. A summary of them is given in the study of Antanasijevic et al. (2014) and these normalization formulas are presented as below:

- The Z-Score transformation:

$$X_s = \frac{X_0 - X_{mean}}{X_{std}} \quad \text{(Eq. 2.17)}$$

- The Sigmoid transformation:

$$X_s = \frac{1}{1+e^{-\frac{X_0 - X_{mean}}{X_{std}}}} \quad \text{(Eq. 2.18)}$$
• The Tanh transformation:

\[ X_s = tanh\left(\frac{x_o - x_{mean}}{x_{std}}\right) \quad (Eq. 2.19) \]

• The Median transformation:

\[ X_s = \frac{x_o}{x_{median}} \quad (Eq. 2.20) \]

where in the above equations (Equation 2.17-2.20) the term \( X_s \) is symbolizing the transformed observations of the parameter \( X \), the term \( X_0 \) is symbolizing the raw observations of the parameter \( X \), the term \( X_{\text{min}} \) is symbolizing the minimum value of the parameter \( X \), the term \( X_{\text{max}} \) is symbolizing the maximum value of the parameter \( X \). Also the \( X_{\text{mean}} \) is symbolizing the mean value of the parameter \( X \), the term \( X_{\text{std}} \) is symbolizing the standard deviation of the parameter \( X \) and the term \( X_{\text{median}} \) is symbolizing the median value of the parameter \( X \).

As stated above PCA is a widely used method to select the most relevant input variables for the ANN model. However, the PCA method is not unique and the selection of the ANN’s input variables can be also done with the method of sensitivity analysis. As stated by Dogan et al. (2009) a sensitivity analysis is used in order to find out key input variables of the ANN and unnecessary input variables must be avoided, because the ANN’s training process might be confused. The basic idea behind the use of sensitivity analysis usage in order to eliminate unnecessary input variables is that those input variables, which might have a weak impact on the ANN’s simulated output, can be excluded from the input variables set (Rankovic et al., 2010). The workflow behind this tactic is to have the input variables to be sequentially excluded one by one by creating different ANNs that are not using them as inputs.
(Cho et al., 2014), and by that way a more compact final ANN model is created (Dogan et al., 2009).

As stated by Chen et al. (2006) this procedure is resembling a neural network trimming process, that it starts from the most complicated neural network (the one with the most input variables) and then is reducing towards the dominant features. The procedure of an ANN trimming is presented as in Table 2.1, where the Chl-a levels of one week ahead (Chl-a(t+dt)) are simulated based on the measured parameters Chl-a, pH, dissolved oxygen (DO), total inorganic nitrogen (TIN), total inorganic phosphorus (TIP), water temperature (Tw), rainfall (R). After the trimming procedure the resulted ANN model with the best performance is selected for the modelling process.

Table 2.1. Procedure of trimming the ANN’s input variables: chlorophyll-a (Chl-a); dissolved oxygen (DO); total inorganic nitrogen (TIN); (TIN), total inorganic phosphorus (TIP); water temperature (Tw); rainfall (R) (taken by Chen et al., 2006)

| Scenarios | Input variables            |
|-----------|-----------------------------|
| S1        | TIN, TIP, Tw, pH, R, Chl-a, DO |
| S2        | TIN, TIP, Tw, pH, R, Chl-a   |
| S3        | TIP, Tw, R, Chl-a            |
| S4        | TIN, Tw, R, Chl-a            |
| S5        | TIN, TIP, R, Chl-a           |
| S6        | TIN, TIP, Tw, Chl-a          |
| S7        | TIN, TIP, Tw, R              |
| S8        | Tw, R, Chl-a                 |

62
2.5. Evaluation of the ANN’s Performance

The performance of an ANN is a very important task of the modelling process, since based on the ANN’s performance the modelling results can be evaluated and the created ANN can be characterized as a reliable/good predictor or not (Hadjisolomou et al., 2016). Also, among a group of created ANNs - which are simulating a specific phenomenon like Chl-α values or dissolved oxygen concentration- the ANN with the highest/better performance is selected to be the most suitable for modelling that specific phenomenon (e.g. Lee et al., 2003; Najah et al., 2011; Sener et al., 2012; Salami Shahid and Ehteshami, 2016). As stated by Gazzaz et al. (2012) among these created ANN models, the one having the highest $R^2$ and the lowest $MSE$ and $RE$ is preferred, while the higher the $R^2$ value and closer to 1 the better.

For the purpose of evaluation of the created ANN’s performance several statistical criteria exist. These criteria are described by the equations (Equation 2.21-2.32), which are presented and explained as below (Nash and Sutcliffe, 1970; Ross, 1996; Dogan et al., 2009; Basant et al., 2010; Liu and Chen, 2012; Antanasijevic et al., 2014; Ahmed, 2014; Sarkar and Pandey, 2015; Cui et al., 2016; Hadjisolomou et al., 2016):

- Root mean square error ($RMSE$): it represents the error associated with the ANN model and it describes an average measure of the error of the simulated output. It is given based on the below equation.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(o_i-s_i)^2}{n}} \quad (Eq. 2.21)$$
- Coefficient of correlation ($R$): it describes the degree of collinearity between the simulated and observed values. If $R=1$ then there is perfect positive linear relationship between the measured and simulated data, while if $R=-1$ then there is perfect negative linear relationship between the measured and simulated data. It is given based on the below mathematical formula.

$$R = \frac{\sum_{i=1}^{n}(o_i-\overline{o})(s_i-\overline{s})}{\sqrt{\sum_{i=1}^{n}(o_i-\overline{o})^2(s_i-\overline{s})^2}} \quad (Eq. 2.22)$$

- Coefficient of determination ($R^2$): it is the square of the correlation coefficient. It is giving the percentage of variability that can be explained by the ANN model and is having the two alternatives mathematical forms as below.

$$R^2 = \left[\frac{\sum_{i=1}^{n}(o_i-\overline{o})(s_i-\overline{s})}{\sqrt{\sum_{i=1}^{n}(o_i-\overline{o})^2(s_i-\overline{s})^2}}\right]^2 \quad (Eq. 2.23)$$

or else the $R^2$ can be written in the alternative form

$$R^2 = \frac{\sum_{i=1}^{n}(o_i-\overline{o})^2-\sum_{i=1}^{n}(o_i-s_i)^2}{\sqrt{\sum_{i=1}^{n}(o_i-\overline{o})^2}} \quad (Eq. 2.24)$$

- Absolute relative error ($RE$): it measures how similar are the predicted values with the measured values. A $RE$ close to zero means that the ANN model had perfect output results.
\[ RE = \frac{1}{n} \sum_{i=1}^{n} \frac{|s_i - o_i|}{o_i} \]  
(Eq. 2.25)

- Average value of residuals (or the bias): it finds out if the ANN model overestimated or underestimated the simulated output value. It is written in the mathematical form given below.

\[ bias = \frac{1}{n} \sum_{i=1}^{n} (s_i - o_i) \]  
(Eq. 2.26)

- Standard error of prediction (SEP): it is given based on the following mathematical equation.

\[ SEP = \sqrt{\frac{\sum_{i=1}^{n} (o_i - s_i - bias)^2}{n-1}} \]  
(Eq. 2.27)

- The Nash-Sutcliffe coefficient of efficiency \( (E_f) \): it compares the \( MSE \) of the ANN model with the variance of the simulated output sequence. When \( E_f \) equals to one, then there is absolute match between the predicted and measured variables. When \( E_f \) equals with negative values (less than zero), then it suggests a not reliable ANN model.

\[ E_f = 1 - \frac{\sum_{i=1}^{n} (o_i - s_i)^2}{\sum_{i=1}^{n} (o_i - \bar{o})^2} \]  
(Eq. 2.28)

- The accuracy factor \( (A_f) \): a factor that is showing the spread of the ANN’s model results. When \( A_f \) equals to one, then there is absolute match between the predicted and measured variables, while when \( A_f \) is having large value then it
means that the ANN does not have good/reliable results. It is given based on the following mathematical formula.

\[ A_f = 10 \left( \frac{\sum \log \left( \frac{s_i}{\sigma_i} \right)}{n} \right) \]  
(Eq. 2.29)

- The index of agreement (IA): It is expressed based on the following mathematical formula.

\[ IA = 1 - \frac{(s_i - o_i)^2}{(s_i - \bar{o})^2 + (o_i - \bar{o})^2} \]  
(Eq. 2.30)

where for the above set of equations the parameters \( o_i \) is the observed value; \( s_i \) the simulated value; \( n \) the observations number; the term \( \bar{o} \) is describing the average observed value calculated as below:

\[ \bar{o} = \frac{1}{n} \sum_{i=1}^{n} o_i \]  
(Eq. 2.31)

the term \( \bar{s} \) is describing the average simulated value calculated as below:

\[ \bar{s} = \frac{1}{n} \sum_{i=1}^{n} s_i \]  
(Eq. 2.32)
2.6. Sensitivity Analysis

Sensitivity analysis is determining/measuring how much sensitive is a model to the changes in the values of the parameters that are used as model’s inputs and to the changes in the structure of the model (Shojaeefard et al., 2013). Sensitivity analysis is a tool of great importance for modelling with ANN and provides useful information about the parameter behavior/interactions. According to Tang et al. (2015) “Sensitivity analysis ascertains how the various sources of uncertainty contribute to the model output uncertainty and system performance, and serves to afford a measure of parameter sensitivity”. Sensitivity analysis is initially investigated, following the realization that an ANN is having perturbation of the output resulted by machine imprecision and noisy input (Yeung et al., 2010). The ANN’s sensitivity analysis can be examined based on several methodologies. The three most well-known sensitivity analysis methods found in the literature are the ‘Perturb’ method, the ‘Weights’ method and the Partial Derivatives method. Meanwhile Shojaeefard et al. (2013) are mentioning the ‘backward stepwise’ method as another very important sensitivity method.

The ‘Perturb’ method is computing the perturbation effect of the input variables regarding the output variable (Dedecker et al., 2005). The effect that a small change of an input variable has on the ANN's output is examined, so the input variables can be classified by an order of importance (Gevrey et al., 2003). The mathematical formula that is describing the ‘Perturb’ method of sensitivity analysis is explained by Lee et al. (2003) as follow:

\[
\text{Sensitivity (%)} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left( \frac{\text{change in output} \%}{\text{change in input} \%} \right)_i \times 100
\]  

(Eq. 2.33)
where the parameter $N_p$ is representing the number of patterns (samples number), constructed with inputs and corresponding outputs for the training set.

The ‘Weights’ method is related with the interpretation of the connections weights from the input layer to the hidden layer (Muttil and Chau, 2007). The contribution of each input parameter to the ANN’s output is measured with the relative parameter importance (I) equation and is given by Lee et al. (2003) as bellow:

$$I = \frac{\sum_{i=1}^{nT} \sum_{j=1}^{nH} |(wij)|}{\sum_{k=1}^{nV} (\sum_{l=1}^{nT} \sum_{j=1}^{nH} |(wij)|)_k} \quad \text{(Eq. 2.34)}$$

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

The ‘PaD’ method or the Partial Derivatives method was firstly introduced by Dimopoulos et al. (1995). This method is following two steps. First the partial derivatives of the output for small changes of each input are computed and afterwards the classification of the relative contributions of each input variable (Gevrey et al., 2003). The sensitivity of the ANN's output for the input variable $x_i$ is symbolized as $SSD_i$ and according to Dimopoulos et al. (1995) is the sum of the squared partial derivatives obtained per input variable and has the following mathematical formula:

$$SSD_i = \sum_{j=1}^{N} (d_{ji})^2 \quad \text{(Eq. 2.35)}$$

where $N$ is the observations number of the data set; $d_{ji}$ the partial derivatives of the output $y_j$ with respect to input $x_j$. The equation for $d_{ji}$ is given bellow:

$$d_{ji} = \sum_{h=1}^{nH} w_{ho} \ast I_{hj} \ast (1 - I_{hj}) \ast w_{ih} \quad \text{(Eq. 2.36)}$$
where \( n_h \) the neurons number of the hidden layer; \( I_{hj} \) the output of the \( h \)-th hidden neuron for the \( j \)-th input; \( w_{ih} \) the weight connecting the \( i \)-th input neuron and the \( h \)-th hidden neuron; \( w_{ho} \) the weight connecting the output and the \( h \)-th hidden neuron; \( S_j \) is the derivative of the output \( y_j \) with respect to input \( x_j \).

### 2.7. Two-way sensitivity analysis

The 'PaD2' method is examining the synergistic effect between the input parameters and is considered as a two-way interaction sensitivity analysis method. The pair-wise interactions of the variables are examined based on the 'PaD2' method. The application of the 'PaD2' method is very similar with the application of the 'Pad' method, but the 'PaD2' method is considered as more advanced modelling, because it considers the two-way interactions between variables, since ecological phenomena are rarely linear or a simple cause originated (Gevrey et al., 2006). As stated by Olay-Marín et al. (2016) the 'Pad' method is making use only of one independent variable at a time, while the 'PaD2' method is calculating the relationship when one predictive variable interacts with another. According to Gevrey et al. (2006) the first part of the 'PaD2' method that gives the partial derivatives of the output \( y_j \) with respect to inputs \( x_{ji} \) and \( x_{j2} \) is given by the following equation:

\[
d_{ji} = S_j \left[ s_j \sum_{h=1}^{n_h} w_{2h} w_{ho} I_{hj} (1 - I_{hj}) \sum_{h=1}^{n_h} w_{2h} w_{ho} I_{hj} (1 - I_{hj}) \right. \\
+ \sum_{h=1}^{n_h} w_{1h} w_{2h} w_{ho} I_{hj} (1 - I_{hj})(1 - 2I_{hj}) \left. \right]
\]

(Eq. 2.37)
where the symbols were explained above, except $s_j$ that is the second derivative of the output neuron with respect to its input (Dimopoulos et al., 1995; van Maanen et al., 2010). The relative contribution of the coupled variables to the ANN is the sum of squared partial derivatives of the coupled variables and is calculated as below:

$$SSD_{12} = \sum_{j=1}^{N}(d_{j12})^2 \quad \text{(Eq. 2.38)}$$

The larger the value of the variable $SSD_i$, the larger impact (influence) the input variable $x_i$ has on the simulated output (van Maanen et al., 2010).

2.8. References

Ahmed, A.A.M. Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). *Journal of King Saud University- Engineering Sciences* 2014, 29, 151-158.

An, Y.; Zou, Z.; Zhao, Y. Forecasting of dissolved oxygen in the Guanting reservoir using an optimized NGBM (1,1) model. *Journal of Environmental Sciences* 2015, 29, 158-164.

Antanasijevic, D.; Pocajt, V.; Peric-Grujic, A.; Ristic, M. Modelling of dissolved oxygen in the Danube River using artificial neural networks and Monte Carlo Simulation uncertainty analysis. *Journal of Hydrology* 2014, 519, 1895-1907.

Areerachakul, S.; Sophatsathit, P.; Lursinsap, C. Integration of unsupervised and supervised neural networks to predict dissolved oxygen concentrations in canals. *Ecological Modelling* 2013, 261-262, 1-7.
Ay, M.; Kisi, O. Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. *Journal of Hydrology* **2014**, *511*, 279-289.

Ball, A. S; De la Rosa, D. Modelling Possibilities for the Assessment of Soil Systems. In *Biological Approaches to Sustainable Soil Systems*; Uphoff, N.; Ball, A. S.; Fernandes, E.; Herren, H.; Husson, M. et al. eds; CRC Press: Boca Raton, FL, USA, 2006, pp. 683-692.

Basant, N.; Gupta, S.; Malik, A.; Singh, K.P. Linear and nonlinear modeling of simultaneous prediction of dissolved oxygen and biochemical oxygen demand of the surface water – A case study. *Chemometrics and Intelligence Laboratory Systems* **2010**, *104*, 172-180.

Ceccaroni, L.; Velickovski, F.; Blaas, M.; Wernand, M.R.; Blauw, A.; Subirats, L. Artificial Intelligence and Earth Observation to Explore Water Quality in the Wadden Sea. In *Earth Observation Open Science and Innovation*, Mathieu PP., Aubrecht C. Eds; Springer: Berlin, Germany, 2018, volume 15, pp. 311-320.

Chau, K.-W. A review on integration of artificial intelligence into water quality modelling. *Marine Pollution Bulletin* **2006**, *52*, 726-733.

Chaves, P.; Tsukatani, T.; Kojiri, T. Operation of storage reservoir for water quality by using optimization and artificial intelligence techniques. *Mathematics and Computers in Simulation* 2004, *67*, 419-432.

Chen, Q.; Morales-Chaves, Y.; Li, H.; Mynett, A.H. Hydroinformatics techniques in eco-environmental modelling and management. *Journal of Hydroinformatics* **2006**, *8*, 297-316, doi:10.2166/hydro.2006.011

Chen, S.H.; Jakeman, A.J.; Norton, J.P. Artificial Intelligence techniques: An introduction to their use for modelling environmental systems. *Mathematics and Computers in Simulation* **2008**, *78*, 379-400.
Cheng, B.; Titterington, D.M. Neural networks: a review from a statistical perspective. 

*Statistical Science* **1994, 9**, 2-30.

Cho, S.; Lim, B.; Jung, J.; Kim, S.; Chae, H. et al. Factors affecting algal bloom in man-made lake and prediction using an artificial neural network. *Measurements* 2014, **53**, 224-233.

Cui, Y.; Zhu, G.; Li, H.; Luo, L.; Cheng, X.; Jin, Y.; Trolle, D. Modelling the response of phytoplankton to reduced external nutrient load in a subtropical Chinese reservoir using DYRESM-CAEDYM. *Lake and Reservoir Management* **2016, 32**, 146–157.

Cybenko, G. Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems* **1989, 2**, 303-314.

Dawson, C.W.; Wilby, R.L. Hydrological modelling using artificial neural networks. *Progress in Physical Geography* **2001, 25**, 80-108.

Dedecker, A.P.; Goethals, P.L.M.; Gabriels, W; De Pauw, N. Optimization of Artificial Neural Network (ANN) model design for prediction of macroinvertebrates in the Zwalm river basin (Flanders, Belgium). *Ecological Modelling* **2004, 174**, 161–173.

Dedecker, A.; Goethals, P.; D'Heygere, T.; Gevrey, M.; Lek, S.; De Pauw, N. Application of Artificial Neural Network models to analyse the relationships between *Gammarus Pulex* L. (Crustacea, Amphipoda) and river characteristics. *Environmental Monitoring and Assessment* **2005, 111**, 223-241.

Demuth, H; Beale, M. Neural Network Toolbox For Use With MaTlab. Version 4, MathWorks Inc.: Natick, MA, USA, 2002.

DeWeber, J.T.; Wagner, T. A regional neural network ensemble for predicting mean daily river water temperature. *Journal of Hydrology* **2014, 517**, 187-200.

Dimopoulos, I.; Bourret, P.; Lek, S. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Processing Letters* **1995, 2**, 1–4.
Ding, Y.R.; Cai, Y.J.; Sun, P.D.; Chen, B. The Use of Combined Neural Networks and Genetic Algorithms for Prediction of River Water Quality. *Journal of Applied Research and Technology* **2014**, *12*, 493-499.

Dogan, E.; Sengorur, B.; Koklu, R. Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *Journal of Environmental Management* **2009**, *90*, 1229-1235.

Emamgholizadeh, S.; Kashi, H.; Marofpoor, I.; Zalaghi, E. Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. *International Journal of Environmental Science and Technology* **2014**, *11*, 645-656.

Engelbrecht, A.P. *Computational Intelligence*. Wiley: England, U.K; 2007

Faruk, D.O. A hybrid neural network and ARIMA model for water quality time series prediction. *Engineering Applications of Artificial Intelligence* **2010**, *23*, 586-594.

Garrett, J.H. Where and Why Artificial Neural Networks Are Applicable in Civil Engineering. *Journal of Computing in Civil Engineering* **1994**, *8*, 129-130.

Gazzaz, N.M.; Yusoff, M.K.; Aris, A.Z.; Juahir, H.; Ramli, M.F. Artificial neural network modelling of the river water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine Pollution Bulletin* **2012**, *64*, 2409-2420.

Gevrey, M.; Dimopoulos, I.; Lek, S. Two-way interaction of input variables in the sensitivity analysis of neural network models. *Ecological Modelling* **2006**, *195*, 43-50.

Gevrey, M.; Dimopoulos, I.; Sovan, L. Review and comparison of methods to study the contribution of variables in artificial neural networks models. *Ecological Modelling* **2003**, *160*, 249-264.

Ghalkhani, H.; Golian, S.; Saghafian, B.; Farokhnia, A.; Shamseldin, A. Real-time flood routing by artificial intelligent models. *Water and Environment* **2013**, *27*, 535-548.
Goethals, P.L.M; Dedecker, A.P.; Gabriels, W.; Lek, S.; De Pauw, N. Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquatic Ecology* 2007, 41, 491-508.

Gupta, M.M.; Jin, L; Homma, N. *Static and Dynamic Neural Networks*. Wiley: New Jersey, USA, 2003.

Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G. Papastergiadou, E. Assessing the Contribution of the Environmental Parameters to Eutrophication with the Use of the “PaD” and “PaD2” Methods in a Hypereutrophic Lake. *International Journal of Environmental Research and Public Health* 2016, 13(8):764, doi: 10.3390/ijerph13080764

Hagan, M.T.; Demuth, H.B.; Beale, M. *Neural Network Design*. PWS: Boston, USA, 1996.

Han, H.-G.; Qiao, J.-F. Adaptive dissolved oxygen control based on dynamics structure neural network. *Applied Soft Computing* 2011, 11, 3812-3820.

Han, H.-G.; Qiao, J.-F.; Chen, Q.-L. Model predictive control of dissolved oxygen concentration based on self-organizing RBF neural network. *Control Engineering Practice* 2012, 20, 465-476.

Haykin, S. Neural Networks, a comprehensive Foundation. Prentice Hall: Upper Saddle River, NJ, USA, 1999.

Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* 1989, 2, 359-366.

Hsu, K.-L.; Gupta, H.V.; Sorooshian, S. Artificial neural network modelling of the rainfall-runoff process. *Water Resources Research* 1995, 31, 2517-2530.

Hu, Y.H.; Hwang, J.N. Introduction to Neural Networks for Signal Processing. In *Handbook of Neural Network Signal Processing*, Hu, Y.H.; Hwang, J.N Eds. CRC Press: Florida, USA, 2002, pp: 12-41.
Kalteh, A.M.; Hjorth, Berndtsson, R. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. *Environmental Modelling & Software* **2008**, *23*, 835-845.

Karul, C.; Soyupak, S.; Çilesiz, A.F.; Akbay, N.; Germen, E. Case studies on the use of neural networks in eutrophication modeling. *Ecological Modelling* **2000**, *134*, 145–152.

Karunanithi, N.; Grenney, W.J.; Whitley, D.; Bovee, K. Neural Networks for River Flow Prediction. *Journal of Computing in Civil Engineering* **1994**, *8*, 201-220.

Kasabov, N.K. *Foundations of neural networks, fuzzy systems, and knowledge engineering*. MIT Press: Massachusetts, USA, 1998.

Kılıç, H.; Soyupak, S.; Gurbuz, H.; Kıvrak, E. Automata networks as preprocessing technique of artificial neural network in estimating primary production and dominating phytoplankton levels in a reservoir: An experimental work. *Ecological Informatics* **2006**, *1*, 431-439.

Kisi, O. Daily River Flow Forecasting Using Artificial Neural Networks and Auto-Regressive Models. *Turkish Journal of Engineering and Environmental Sciences* **2005**, *29*, 9-20.

Kmet, T.; Kmetova, M. Adaptive critic design and Hopfield neural network based simulation of time delayed photosynthetic production and prey-predator model. *Information Sciences* **2015**, *294*, 586-599.

Kohonen, T. The Self-Organizing Map. *Proceedings of the IEEE* **1990**, *78*, 1464-1480.

Kruszewski, A.; Wang, R.; Guerra, T.M. Nonquadratic stabilization conditions for a class of uncertain discrete time TS fuzzy models: a new approach. *IEEE Transactions on Automatic Control* **2008**, *53*, 606-611.

Lee, J.H.W.; Huang, Y.; Dickman, M.; Jayawardena, A.W. Neural networking modelling of coastal algal blooms. *Ecological Modelling* **2003**, *159*, 179-201.
Lee, S.; Park, I.; Koo, B.J.; Ryu, J.-H.; Choi, J.-K.; Woo, H.J. Macrobenthos habitat potential mapping using GIS-based artificial neural network models. *Marine Pollution Bulletin 2013, 67*, 177-186.

Lek, S.; Guegan, J.F. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling 1999, 120*, 65-73.

Li, H.; Hou, G.; Dakui, F.; Xiao, B.; Song, L.; Liu, Y. Prediction and elucidation of the population dynamics of *Microcystis* spp. in Lake Dianchi (China) by means of artificial neural networks. *Ecological Informatics 2007, 2*, 184–192.

Lippmann, R.P. *An introduction to computing with neural nets. IEEE ASSP Magazine 1987, 4*, 4-22.

Liu, W.-C.; Chen, W.-B. *Prediction of water temperature in a subtropical subalpine lake using an artificial neural network and three-dimensional circulation models. Computers and Geosciences 2012, 45*, 13-25.

Maier, H. R.; Dandy, G. C; Burch, M.D. Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. *Ecological Modelling 1998, 105*, 257-272.

Maier, H.R.; Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. *Environmental Modelling & Software 2000, 15*, 101-124.

Millie, D.F.; Weckman, G.R.; Young II, W.A.; Ivey, J.E.; Carrick, H.J.; Fahnenstiel, G.L. Modeling microalgal abundance with artificial neural networks: Demonstration of a heuristic ‘Grey-Box’ to deconvolve and quantify environmental influences. *Environmental Modelling & Software 2012, 38*, 27-39.

Muttil, N.; Chau, K.W. Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Applications of Artificial Intelligence 2007, 20*, 735–744.
Najah, A.; El-Shafie, A.; Karim, O.A.; Jaafar, O. Integrated versus isolated scenario for predicting dissolved oxygen at progression of water quality monitoring stations. *Hydrology and Earth System Sciences* 2011, 15, 2693-2708.

Nash, J.E.; Sutcliffe, I.V. Riverflow forecasting through conceptual models. *Journal of Hydrology* 1970, 273, 282-290.

Olaya-Marín, E.J.; Martínez-Capel, F.; García-Bartual, R.; Vezza, P. Modelling critical factors affecting the distribution of the vulnerable endemic Eastern Iberian barbel (*Luciobarbus guiraonis*) in Mediterranean rivers. *Mediterranean Marine Science* 2016, 17, 264-279.

Ranković, V.; Radulović, J.; Radojević, I.; Ostijic, A.; Comic, L. Neural network modeling of dissolved oxygen in the Gruza reservoir, Serbia. *Ecological Modelling* 2010, 221, 1239-1244.

Recknagel, F., French, M.; Harkonen, P.; Yabunaka, K.I. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling* 1997, 96, 11-28.

Ross, T. Indices for performance evaluation of predictive models in food microbiology. *The Journal of Applied Bacteriology* 1996, 81, 501–508.

Rumelhart, D.E.; Hinton, G.E.; Williams, R.G. Learning representations by backpropagation errors. *Nature* 1986, 323, 533-536.

Salami Shahid, E.; Ehteshami, M. Application of artificial neural networks to estimating DO and salinity in San Joaquin River basin. *Desalination and Water Treatment* 2016, 57, 4888-4897.

Sanchez Mesa, J.A.; Galan, C.; Hervas, C. The use of discriminate analysis and neural networks to forecast the severity of the Poaceae pollen season in a region with a typical Mediterranean climate. *International Journal of Biometeorology* 2005, 49, 355-362.
Sarkar, A.; Pandey, P. River Water Quality Modelling using Artificial Neural Network Technique. *Aquatic Procedia* 2015, 4, 1070-1077.

Scardi, M.; Harding, L.W., Jr. Developing an empirical model of phytoplankton primary production: A neural network case study. *Ecological Modelling* 1999, 120, 213–223.

Sener, E.; Terzi, O.; Sener, S.; Kucukkara, R. Modeling of Water Temperature Based on GIS and ANN Techniques: Case Study of Lake Egirdir (Turkey). *Ekoloji* 2012, 21, 83, 44-52.

Seo, Y.; Kim, S.; Kisi, O.; Singh, V. P. Daily water level forecasting using wavelet decomposition and artificial intelligence techniques. *Journal of Hydrology* 2015, 520, 224-243.

Shang, Z. Application of Artificial Intelligence CFD based on neural network in vapor-water two-phase flow. *Engineering Applications of Artificial Intelligence* 2005, 18, 663-671.

Shojaeefard, M. H.; Akbari, M.; Tahani, M.; Farhani, F. Sensitivity Analysis of the Artificial Neural Network outputs in Friction Stir Lap Joining of Aluminum to Brass. *Advances in Materials Science and Engineering* 2013, doi.org/10.1155/2013/574914.

Simon, H. Neural Networks: A Comprehensive Foundation. Pearson Prentice Hall: Delhi, India, 2005.

Srinivasan, D.; Liew, A.C.; Chang, C.S. A neural network short-term load forecaster. *Electric Power Systems Research* 1994, 28, 227-234.

Tang, S.; Sun, T.; Shen, X.; Qi, M.; Feng, M. Modeling net ecosystem metabolism influenced by artificial hydrological regulation: An application to the Yellow River Estuary, China. *Ecological Engineering* 2015, 76, 84-94.

Tomenko, V.; Ahmed, S.; Popov, V. Modelling constructed wetland treatment system performance. *Ecological Modelling* 2007, 205, 355-364.
Turan, E.M.; Yurdusev, A.M. River flow estimation from upstream flow records by artificial intelligence methods. *Journal of Hydrology* 2009, *369*, 71-77.

Van Maanen, B.; Coco, G.; Bryan, K. R.; Ruessink, B. G. The use of artificial neural networks to analyze and predict alongshore sediment transport. *Nonlinear Process in Geophysics* **2010**, *17*, 395-404.

Veelenturf, L.P.J. *Analysis and Applications of Artificial Neural Networks*. Prentice Hall: Hertfordshire, U.K, 1995.

Vilas, L.G.; Spyrakos, E; Palenzuela, J.M.T. Neural network estimation of chlorophyll a from MERIS full resolution data for the coastal waters of Galician rias (NW Spain). *Remote Sensing of Environment* **2011**, *115*, 524–535.

Wang, R.; Sun, H.; Zha, B.; Wang, L. Research and application of Improved AGP algorithm for Structural Optimization Based on Feedforward Neural networks. *Mathematical problems in Engineering* **2015**, doi: 10.1155/2015/481919.

Wang, W.-C.; Chau, K.-W.; Cheng, C.-T.; Qiu, L. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology* 2009, *374*, 294-306.

Wu, N.; Huang, J.; Schmalz, B.; Fohrer, N. Modeling daily chlorophyll a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches. *Limnology* 2014, *15*, 47-56.

Yeung, D., S., Cloete, I.; Shi, D.; Ng, W. W.Y. *Sensitivity Analysis for Neural Networks*. Springer-Verlag: Heidelberg, Germany, 2010.

Zaji, A.M.; Bonakdari, H. Robustness lake water level prediction using the search heuristic-based artificial intelligence methods. *ISH Journal of Hydraulic Engineering* 2018, doi.org/10.1080/09715010.2018.1424568
Zhang, J.; Ding, W. Prediction of Air Pollutants Concentration based on Extreme Learning Machine: The case of Hong Kong. *International Journal of Environmental Research and Public Health* 2017, 14, 114; doi:10.3390/ijerph14020114.

Zhang, P.; Hong, B.; He, L.; Cheng, F.; Zhao, P.; Wei, C.; Liu, Y. Temporal and spatial simulation of atmospheric pollutant PM$_{2.5}$ changes and risk assessment on population exposure to pollution using optimization algorithms of the back propagation-Artificial Neural Network model and GIS. *International Journal of Environmental Research and Public Health* 2015, 12, 12171–12195.

Zounemat-Kermani, M. Principal Component Analysis (PCA) for estimating chlorophyll concentration using forward and generalized regression neural networks. *Applied Artificial Intelligence* 2014, 28, 16–29.
3. Unsupervised Neural Networks.

3.1. Self-Organizing Map (SOM) neural networks

The Kohonen's Self Organizing Map (SOM) is a special category of ANNs and more details of its origin can be found in the studies of Kohonen (1982); Kohonen (1998); Kohonen (2001). SOM is an unsupervised learning algorithm of ANNs, meaning that no human intervention is required during the learning process of the ANN (Zhang et al., 2008). The term unsupervised is attributed to SOM because no supervision is needed for its learning process. The term self-organizing is derived from the SOM's ability to learn and organize information without being given the associated output values for the corresponding input data, while the desired output is not known a priori (Al Mudhaf et al., 2010).

SOMs are given great concern as modelling tools in water quality studies, since they have the ability to analyze multivariant data by means of a sophisticated visualization capacity and they are a good alternative to the classical water quality modelling techniques (Cinar and Merdun, 2009). Generally, the SOM models are considered as a superior modelling technique comparing with the classical models used in water quality studies (Hadjisolomou et al., 2018), something that is discussed in more detail in a following chapter (see Chapter 4).

The SOM neural networks have started to be used widely in the water quality modelling because they are cable to resolve some drawbacks that other ANNs are facing, like the satisfactory description of the cause-effect relationship between multivariate data sets (Cinar and Merdun, 2009). Some examples of hydrological studies that are using the SOM neural Network are presented next, in order to highlight the significance of the SOM method in water quality modelling. In their study of Li et al. (2017) clustered 27 lakes in China into four groups with the use of a SOM and associated the 24 monitored water quality parameters
with a trophic level index, in order to evaluate how prone was each lake to eutrophication. In another study of Lee and Scholz (2006a) a SOM neural network was used in order to predict the heavy metal removal performance in experimental constructed wetlands and to characterize the heavy metal removal mechanisms. In their study Al Mudhaf et al. (2010) used the SOM neural network, aiming to examine if the levels of trihalomethanes (disinfection byproducts of health concern) in desalinated drinking water depended on the geographical location of water intake.

The SOM neural network is having many modelling abilities, which are making it an attractive modelling technique in water quality sciences. Several of these abilities are presented below as mentioned by Buscema (2010):

1. The SOM processes all the records and all the variables simultaneously.
2. SOMs are not sensitive to variables cardinality.
3. The SOM neural network can process non-linear relationships among the data.
4. The SOM clusters the whole data according to global similarities.
5. The SOM projects multidimensional data into a two-dimensional space and during the multidimensional scaling it selects only the most important features of the data set.
6. The SOM visualizes the results with figures in the form of tables and maps that are easy to be implemented.

In many SOM water quality modelling studies (e.g. Kangur et al., 2007; Jahan et al., 2013; An et al., 2016), the visualization ability and the dimension reduction ability of the SOM are emphasized as very important advantages of the SOM neural network. Regarding the SOM's visualization ability Cho et al. (2009) stated that an advantage of the SOM is its
efficiency to visualize large amounts of data through a component plane technique. In their study Jahan et al. (2013) stated that one of the SOM's most significant characteristics is that the results obtained from the SOM can be visualized from its topology map. The SOM can interpret information from multiple variables with its visualization abilities (Juntunen et al., 2013). Also from the SOM's map visual inspection it is provided simultaneous observation of both spatial and temporal changes in water quality (Astel et al., 2007).

Regarding the dimension reduction ability of the SOM Peeters et al. (2007) stated that the SOM is projecting multidimensional data into a two-dimensional grid in a topology preserving way and is associating non-linear complex relationships between the variables. The SOM is having the ability to represent the information of multi-dimensional environmental data into fewer dimensions (Chon, 2011). For the convenience of visualization, the artificial neurons (nodes) of the SOM are arranged in a limited number of dimensions, preferably two dimensional grids of neurons (Kohonen, 1982). Each neuron is associated with $m=[m_1, ..., m_i]$ that is an $i$-dimensional weight vector (or prototype vector), where parameter $i$ equals to the dimension of the input vectors (see Figure 3.1), and the neurons are connected to adjacent neurons by a neighborhood relation based on which the topology of the map is dictated (Vesanto et al., 2000).
Figure 3.1. Representation of dimension reduction from a \( i \)-dimensional space (where \( i \) the number of input parameters) to a 2-dimensional space with the use of SOM. The measured data are rearranged to a 2D array in order to have a row vector. The resulted row vector is used to update the weight of the SOM via an unsupervised learning algorithm (taken by Liu and Weisberg, 2011).

3.2. The structure and basic steps of the SOM algorithm

The SOM topology is formed based on the local lattice structure and the global map shape. The local lattice structure can be rectangular or hexagonal and the centermost unit is associated with several discrete neighborhoods (size 0, 1 and 2), where the first inner polygon is giving the 0-neighborhood, the second inner polygon the 1-neighborhood and the third inner polygon the 2-neighborhood (see Figure 3.2).
Figure 3.2. Representation of the discrete neighborhoods, where the first inner polygon is giving the 0-neighborhood, the second inner polygon the 1-neighborhood and the third inner polygon the 2-neighborhood: (a) hexagonal lattice; (b) rectangular lattice (taken by Vesanto et al., 2000).

On that rectangular or hexagonal grid the neurons of the output layer are arranged, and each of the input layer vectors is connected to each of the neurons of the output layer through a weight vector (Peeters et al., 2007; Choi et al., 2014). The possible map shapes of the SOM are the: a) the sheet shape, which is the most commonly used (characterized as the default), b) the cylinder shape and c) the toroid shape. The different map shapes are visualized as in the following figure (Figure 3.3).
The SOM algorithm is following several steps that are repeated iteratively. A synoptic description of the SOM algorithm’s procedure is given in the studies of Barge and Sharif, 2016: initially the connection weights are initialized with small random values. After the lowest value of the Euclidean distance between the input and weight vectors must be found, based on competition between the output neurons. Then the Best Matching Unit (BMU) is calculated, where as BMU is defined the winning neuron from competition. Finally, the stimulated neurons adjust their weights by minimizing the squared Euclidean distance between the input and weight vectors.

This means that BMU and its neighboring (adjacent) neurons are changing their weights for each iteration, in order to reduce the distance between the input vector and the weights (Kohonen, 2001; Nourani et al., 2013). The learning procedure of the SOM algorithm is analytically explained in the following steps (Kangas and Kohonen, 1996; Kohonen, 2001;
1. From the given data set, which is consisted of \( n \) columns of variables and \( m \) rows of measured samples, a \( mxn \) matrix is created. For avoiding bias in the classification because of different unit’s magnitude of the different variables, the variables are normalized before presenting them to the SOM.

2. Initialize the weight vector with random values, which are ranging within the interval \([0, 1]\). The initial values for the map size, the maximum value of possible iterations \( T \), the neighborhood function \( h_c(0) \), the initial value of learning ratio \( a(0) \) (where \( 0 < a(t) < 1 \)). The learning ratio \( a(t) \) factor is controlling the rate of change of the reference vectors and may have several forms, that are given based on the following mathematical functions (Natita et al., 2016):

- **Linear function:**
  \[
  a(t) = a_0 \cdot \frac{1}{t} 
  \]  
  (Eq. 3.1)

- **Power function:**
  \[
  a(t) = a_0 \cdot e^{t/T} 
  \]  
  (Eq. 3.2)

- **Inverse function:**
  \[
  a(t) = a_0 (1 - \frac{t}{T}) 
  \]  
  (Eq. 3.3)
where in the functions above the term $T$ is the training length and the term $a_0$ is initial learning rate. The schematically representation of the above learning rate functions are given as in Figure 3.4.

**Figure 3.4.** Schematically representation of the learning rate functions, where the linear learning function is drawn with the red line, the power learning function is drawn with the black line and the inverse learning function is drawn with the blue line.

The neighborhood function $h_{ci}(t)$ is defining the neighborhood kernel around the winner unit $c$ (Lee and Scholz, 2006a) and it determines the size and the nature of the neighborhood around the winner neuron (Ramadas et al., 2003). The neighborhood function $h_{ci}(t)$ can be found in the two following forms (which are the most popular ones):

- Bubble (or square) neighborhood function (Ramadas et al., 2003):
where in the above equation the $r_c$ and the $r_i$ are representing the positions of the winner neuron $c$ and the neuron $i$ in the SOM; $\|r_c, r_i\|$ is the distance between the winner neuron $c$ and the neuron $i$; $\sigma(t)$ is the neighborhood radius at time $t$. Therefore, the Bubble function is characterized as a step function, which is constant over the whole neighborhood of the winner neuron and zero elsewhere (Vesanto et al., 2000). The shape of the Bubble neighborhood function is graphically represented in the figure below (Figure 3.5)

![Graphical representation of the Bubble neighborhood function.](image)

**Figure 3.5.** Graphical representation of the Bubble neighborhood function.

- Gaussian neighborhood function (Ramadas et al., 2003):

$$h_{ci}(t) = \begin{cases} 
1 & \text{when } \|r_c, r_i\| < \sigma(t) \\
0 & \text{otherwise}
\end{cases} \quad (\text{Eq. 3.4})$$

$$h_{ci}(t) = \begin{cases} 
\exp\left(-\frac{\|r_c, r_i\|^2}{2\sigma^2(t)}\right) & \text{when } \|r_c, r_i\| < \sigma(t) \\
0 & \text{otherwise}
\end{cases} \quad (\text{Eq. 3.5})$$
where in the above equation the $\sigma(t)$ is the neighborhood radius at time $t$; the $r_c$ and the $r_i$ are representing the positions of the winner neuron $c$ and the neuron $i$ in the SOM. The shape of the Gaussian (or Mexican-hat shaped) neighborhood function is graphically represented in the next figure (Figure 3.6).

![Graphical representation of the Gaussian neighborhood function.](image)

**Figure 3.6.** Graphical representation of the Gaussian neighborhood function.

3. An input vector is presented to the network and then distances between the given input vector and the $i$ weight vector are calculated. The Euclidean distance (symbolized as $D_i$) is the most frequently used and the mathematical equation that describes this relationship is given as below:

$$D_i = \sqrt{\sum_{j=1}^{R} (P_{ij} - W_{ij})^2} \quad ; \quad i = 1, 2, ..., S \quad \text{(Eq. 3.6)}$$
where $S$ is the number of output neurons, $R$ is the dimension of the input vectors, $p_{ij}$ represents the $j$ element of the input vector, and $w_{ij}$ symbolizes the $j$ element of the $i$ weight vector.

4. The smallest distance is chosen and the Best Matching Unit (BMU), that is defined as the neuron with the weight vector closest to the input variable $x$, as given by the equation:

$$
\|x - m_c\| = \min(\|x - m_i\|) \quad \text{(Eq. 3.7)}
$$

where $\| \|$ symbolizes the distance measure, $x$ the input vector, $m$ the weight vector, and $c$ the subscription of the weight vector for the winning neuron.

5. The weights of the BMU and the unit within its neighborhood ratio $N(t)$ are updated in order the new reference vectors to be closer to the input vector (see Figure 3.7). During that process the neighborhood ratio function $h_{ci}(t)$ and the learning ratio $a(t)$ are decreasing along with the number of iterations of the model, forcing by that way the network to converge. The weight vector is updated based on the following relationship:

$$
w_{ij}(t + 1) = w_{ij}(t) + a(t)h_{ci}(t)(p_{ij} - w_{ij}(t)) \quad \text{(Eq. 3.8)}
$$

where $w_{ij}(t + 1)$ is symbolized the weight vector at the learning step $t+1$. 


Figure 3.7. Schematically representation of the updating process of the Best matching unit (BMU) and its neighbors towards the marked with x input sample. The solid lines are associated with situation before updating and the dashed lines are associated with the situation after updating (taken by Vesanto et al., 2000).

6. The process goes on with an interactive way and the steps 3 to 5 are repeated, until a predefined number of iterations is done and then the process starts again from step 3.

After training the SOM its quality must be assessed. The SOM quality is usually measured with the use of quantization error (QE) and the topographic error (TE) (Lee and Scholz 2006a), where the QE shows if the neurons of the trained SOM adapt well to the input variables and the TE shows how well the trained SOM keeps the topography of the analyzed data (Stefanovic and Kurasova, 2011). According to Kohonen (2001); Liu et al. (2006) the QE is defined as the mean distance between each data point and its BMU, while the TE gives the
percentage of the data vectors for which the first BMU and the second BMU are not neighboring units. The mathematical formulas that describe the QE and the TE are given accordingly as below (Stefanovic and Kurasova, 2011; Tsai et al., 2017):

\[ E_{QE} = \frac{1}{m} \sum_{p=1}^{m} \| X_p - M_{c(p)} \| \quad \text{(Eq. 3.9)} \]

where \( E_{QE} \) is the quantization error; \( m \) is the number of input vectors of the SOM; \( X_p \) the data vectors; \( M_{c(p)} \) the winner vectors of \( X_p \).

\[ E_{TE} = \frac{1}{m} \sum_{p=1}^{m} u(X_p) \quad \text{(Eq. 3.10)} \]

where \( m \) is the number of input vectors of the SOM; \( X_p \) is a data vector. The \( u \) function equals to \( u(X_p) = 1 \) if the first and second BMUs of \( X_p \) are not adjacent to each other; otherwise the \( u(X_p) = 0 \).

There are no specific rules for choosing SOM's parameters like global topology type (sheet, cylinder and toroid); local topology type (hexagonal and rectangular); neighbor function, map dimension etc (Al Mudhaf et al. 2010). However, a very commonly used rule (Park et al., 2006; Al Mudhaf et al. 2010) for finding the map dimension is the one proposed by Vesanto and Alhoniemi (2000) and has the following formula:

\[ M = 5 \sqrt{\langle n \rangle} \quad \text{(Eq. 3.11)} \]

Another proposed way for determining the optimum map size of the SOM is to found for which neurons number the quantization error (QE) and the topographic error (TE) are
minimized (Lee and Scholz, 2006a; Jeong et al., 2010; Wang et al., 2014; An et al., 2016). This can be archived by examining several map sizes of the trained SOM and to find for which one the TE and QE are having their lowest values (Wang et al., 2014), where this procedure can be graphically represented in the following figure (Figure 3.8) and the optimum map size is the one with the 3X3 map size for the specific SOM application that is described in the study of Tsai et al. (2017).

**Figure 3.8.** Graphical plots of the quantization error (QE) and the topographic error (TE) in order to find the optimal map size of the trained SOM. The optimal map size is found to be the 3X3 (taken by Tsai et al., 2017).

The TE and QE are decreasing gradually as the map size of the trained SOM is increasing (see Figure 3.9) and the optimum size is based on local minimum values for the parameters TE and QE, however the heuristic rule as suggested by Vesanto and Alhoniemi
that is given in the above equation (Equation 3.11) fits well the water quality data (Cereghino and Park, 2009).

Figure 3.9. Graphically examination of the quantization error’s (QE) and the topographic error’s (TE) behavior for different map sizes. The heuristic rule as suggested by Vesanto and Alhoniemi (2000) that is given in the above equation (Equation 3.11) fits well the ecological data for the point of the 66 units (nodes) (taken by Cereghino and Park, 2009).

The input data before being presented to the SOM network must be pre-processed, because different magnitude of the water quality variables may result to unequal importance for the weight update procedure, and by that way the SOM model performance is improved (Cinar and Merdun, 2009). Furthermore, all the input variables are treated by the SOM in the same way (Garcia and Gonzalez, 2004). In their study Zhang et al. (2009) are emphasizing that the SOM modelling results might be affected by an inappropriate pre-processing method and they propose a way to resolve this problem based on the Shapiro-Wilk normality test in order to find of the measured variables were following normal distributions.
Several data pre-processing (transformation) methods for the SOM input parameters are found in the literature (e.g. Garcia and Gonzalez, 2004; Peeters et al., 2007; Cinar and Merdun, 2009; Zhang et al., 2009, An et al., 2016) and are presented as below:

- The logarithmic normalization:

  \[ x^{tr} = \log(x + \alpha) \quad \text{(Eq. 3.12)} \]

  where \( x^{tr} \) symbolizes the transformed data, \( x \) symbolizes the initial data, \( \alpha \) is a real number.

- The zero-mean (z-score) normalization:

  \[ x^{tr} = (x - x_{\text{mean}})/\sigma_x \quad \text{(Eq. 3.13)} \]

  where \( x^{tr} \) symbolizes the standardized data, \( x \) symbolizes the initial data, \( x_{\text{mean}} \) the mean deviation if \( x \), \( \sigma_x \) the standard deviation of \( x \).

- The Min-Max normalization:

  \[ x^{tr} = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)} \quad \text{(Eq. 3.14)} \]

  where \( x^{tr} \) is the transformed data \( x \), \( \text{min}(x) \) the minimum value of \( x \), \( \text{max}(x) \) the maximum value of \( x \).

There is no particular constrain for transforming the SOM’s input variable data for any of the mentioned normalizations above, since as it stated by Giraudel and Lek (2001) that in
opposition with the correspondence analysis, it is possible to have negative numbers in the data sample and a standardization of the measured parameters can be made before presenting the input variables to the SOM model.

### 3.3. U-matrix and Component Planes

The SOM is visualized with the usage of the U-matrix and the component planes (CPs), and by that way the clustering structure of the SOM map is presented (Dong et al., 2012). The unified distance matrix (U-matrix) or otherwise known as “Interneuron distance matrix” (Garcia and Gonzalez, 2004) is very important. The U-matrix of the trained SOM is visualizing the distances between the SOM map neurons (Vesanto and Alhoniemi, 2000; Lee and Scholz; 2006a). The U-matrix is graphically presented as in the Figure 3.10. Usually the distances between the neighboring SOM map neurons are visualized by using gray shading scales between each neuron (Tota-Maharaj and Scholz, 2013), however this is not a norm and other color shading scales can be used.

The darker the shade the largest is the Euclidean distance between the neurons of the SOM map, therefore a light shade between the neurons of the SOM map is indicating a small Euclidean distance (Vesanto et al., 1999; Recknagel et al., 2006b). The darkest shades on the U-matrix are representing biggest distances between neighboring data and therefore the borders between the clusters are defined (Recknagel et al., 2006a, Farsadnia et al., 2014), and by that way it is possible to achieve the identification and visualization on the U-matrix for the several clusters of the input environmental data (Scholz, 2011).
Figure 3.10. U-matrix of the trained SOM map, where the distances between the neurons are visualized by using the gray scales. The dark shades indicate long distance between the neurons, while the lighter scales indicate shorter distance (taken by Voutilainen et al., 2012).

The CPs are used in order to visualize the value of the SOM’s input variables in each map unit (Vesanto and Alhoniemi, 2000; Lee and Scholz; 2006a). The CPs are considered as a good tool to evaluate the interrelationships between the several water quality parameters (An, et al., 2016; Hadjisalomou et al., 2018). These interrelationships between the SOM input variables can be observed by examining the Figure 3.11, which is representing the CPs of several surface water quality parameters measured in the area of Hong Kong.

Through the CPs the exploratory potential among the SOM’s input variables is provided and therefore the CPs visualization are making the SOM neural network to be characterized as a useful tool in exploratory data analysis and clustering of multivariate data sets (Peeters et al., 2007; Hadjisalomou et al., 2018), while the visualization of inter-specific
association among the CPs is supporting the SOMs role as a useful tool in exploratory data analysis (Giraudel and Lek, 2001). The CPs mapping interpretation is based on degrading color shadings. Each CP is associating the value range of every input variable of the trained SOM with each map unit (Vesanto et al., 1999).

**Figure 3.11.** Visualization of the Component Planes (CPs) associated with the trained SOM that is examining the descriptive characteristics of surface water quality in the area of Hong Kong based on several measured water quality parameters (taken by An et al., 2016).
The SOM is a technique used in data analysis clustering for exploring, in order to discern any reasonable relationships among the environmental variables and in most modelling cases without prior knowledge or assumptions of the given data set (Bieroza et al., 2009). Therefore, the term “correlation hunting” is given by Vesanto (1999) that states that the CPs can also be used for correlation hunting among the SOM’s input variables. The possibility to have even local correlations might exist, when two CPs are resembling each other only to some regions (Barreto and Perez-Uribe, 2007). So, the term correlation is not limited only to just linear correlation, but also includes nonlinear correlation and local (partial) correlation between the SOM’s input variables (Vesanto and Ahola, 1999).

The SOM can identify the corresponding correlations and similarities between the water quality variables (Dong et al., 2012). When there is strong positive correlation between two water quality variables, then their associated CPs are very identical. When there is strong negative correlation between two water quality variables, then their associated CPs are identical but with reverse degrading of color shading (Zhang et al., 2015; Hadjisalomou et al., 2018). For example, in the above figure (Figure 3.11) by examining the CPs of the Total Phosphorus parameter and the Chemical Oxygen Demand parameter it is observed that there is strong positive correlation between these two water quality parameters, since the CPs are very similar and the dark scale coloring (red) areas are following an identical degradation towards the light scale coloring (blue) areas.

The spherical/complete implementation of the SOM results requires some visualization and clustering tools to be used, in order to have substantial information on the input data distribution and interrelationships among the SOM’s input variables (Bieroza et al., 2009). Besides the U-matrix and the CPs, the sample distribution map (see Figure 3.12) and the hit histogram (see Figure 3.13) are also used for implementing the results of the trained SOM. With the use of the sample distribution map for each map node the most frequent BMU with
an assigned data set index (e.g. sampling site number) is associated (Bieroza et al., 2009; Zhang et al., 2015), meaning that with the help of the sample distribution map the sampling sites (or every other possible assigned index) are distributed and clustered analogously.

Figure 3.12. A sample distribution map (right) along with the associated U-matrix. The capital letters are indicating sampling events with time order (e.g. A, B, C…) of dissolved organic matter, the Arabic numbers are for sampling sites of a watershed, while the “a” symbol is for unpolluted and the “b” symbol is for polluted (taken by Zhang et al., 2015).

By that way the sampling sites can be related with the CPs maps and to make associations about the environmental conditions that are characterizing each sampling site, like their environmental parameters value levels (therefore to evaluate the behavior and interactions of the water quality parameters for that specific sampling site), to exam their grouping and similarity/ or dissimilarity (e.g. comparison of mean, maximum, minimum
values) with other sampling sites and try to find possible reasons/events that are involved for this.

Figure 3.13. (a) Hit histogram of the trained SOM, where the number in the cell represents the sample hits for that cell; (b) Hit histogram of the trained SOM, where the SOM analysis separated unpolluted samples of the watershed with red and the polluted samples with green (taken by Zhang et al., 2015).

The hit histogram of the trained SOM is displaying the distribution of sample on the SOM map (Bieroza et al., 2009; Zhang et al., 2015, Hadjisolomou et al., 2018). Specifically, for Figure 3.13 (a) the hit histogram of the trained SOM is displaying the distribution of the number of sample hits for each cell (node). This hit histogram is illustrating how many times each prototype neuron is the winning neuron for the data set of the water quality variables and the neurons with higher number of hits are representing more data samples with similar water quality properties (Zhang et al., 2015). Regarding the Figure3.13 (b), the hit histogram of the trained SOM is displaying the distinction between the data set samples (e.g. polluted or
unpolluted water samples) and the distribution of the density of sample hits for each cell (node), where the more samples hits for each cell the more intense (having larger surface) is the color inside it.

In conclusion it can be said that this visualization of the hit histogram of SOM analysis is displaying the density of the SOM hits based on a distinction of the data sample. According to Bieroza et al., (2009) the greater the spread of water quality samples on the hit histogram, then the greater the variation of the water quality properties that can be observed. This is related with the fact that the water quality samples assigned to distant neurons of the SOM map are different, while these in neighboring neurons are similar (Plociennik et al., 2015).

When referring about the sample distribution, then each neuron of the SOM map is giving the assigned name of a sampling site (or every other possible assigned index) of the most frequent best matching sample and is representing many data samples that are assigned to this winning neuron; while in the case of the hit histogram then is given the number of those data samples that are associated with the winning neuron (Zhang et al., 2015).

### 3.4. The Cluster Structuring Index

The influence of each input variable on the organization of the SOM map is measured with the use of an index, called the Cluster Structuring Index (CSI or SI). The SI was initially created to assess which of the measured diatom species were having the strongest influence on the organization of the SOM map (Park et al., 2005; Park et al., 2006). As stated by Park et al. (2005) the idea behind the SI relied upon the observation that when a cluster defined in the SOM is an assembly of neighboring units of the SOM, and therefore when a variable is specific to a certain cluster and is having smaller difference between the weights of SOM
units. The SI for the $i$-th variable can be calculated based on the formula given below (Park et al.; 2005; Park et al., 2006):

$$SI_i = \sum_{j=1}^{S} \sum_{k=1}^{S-1} \frac{|w_{ij} - w_{jk}|}{||r_j - r_k||}$$  \hspace{1cm} (Eq. 3.15)

where $w_{ij}$ and $w_{ik}$ are symbolized the connection weights for the $i$-th input variable for neurons $j$ and $k$ of the SOM map; $||r_j - r_k||$ is the topological distance between units $j$ and $k$; and $S$ is the total number of SOM output units.

According to Tison et al. (2005); Park et al. (2006) the SI is having high values for variables that are observed in high abundance of frequency in data samples found in the same area of the SOM map, while the SI is having low values for variables observed in many data samples of different clusters. As it is observed by Tsai et al. (2017) the existence of a strong gradient for an input variable generates a high SI value, while the existence of a weak gradient for an input variable generates a low SI value.

The SI is calculated in several water quality modelling studies that are using a SOM neural network (e.g. Tison et al.; 2005; Park et al., 2006; Tsai et al., 2017). For example, in the study of Tsai et al. (2017) a SOM neural network was created in order to explore the multi-relation between fish species and water quality parameters. The eight measured water quality parameters were the dissolved oxygen (DO); total phosphorus (TP); water temperature (WT); pH; electric conductivity (EC); suspended solids (SS); chemical oxygen demand (COD); ammonia nitrogen (NH3-N). The SOM managed to produce meaningful patterns among the 25 fish species and the water quality parameters for the 276 heterogeneous data sets measured across Taiwan, while the SI regarding the water quality parameters was evaluated (for more see Fig. 3.14).
**Figure 3.14.** The structuring index (SI) that is calculating the influence of the monitored water quality parameters on the organization of the SOM map (taken by Tsai et al., 2017).

A problem that might appear during the process of modelling with the use of a SOM neural network is when too many input variables exist. This may result to have difficulties in the evaluation of variable's contribution with the use of SOM's component planes, because of the large number of input variables. In their study Tison et al. (2005) are solving this problem by using the SI in order to reduce the number of input variables. More specifically it was decided to keep only the variables with SI over 90 and by that way the number of input variables was reduced by three, while no significant quantities of ecological information were lost. In order to exam if the ecological information is preserved for the new reduced data set; the regression determination coefficient ($R^2$) between SI values of the new reduced data set and the initial data set is calculated.

In the study of Park et al. (2006) the problem of too many input variables of a SOM is resolved in a similar way, where the SI was used to create several reduced data classes from the original data set and the created classes with small SI were excluded (see Fig. 3.15). Then
the new created data class is chosen based on the examination of the sum of the Euclidean distances of the SI values of the original data set and the reduced data set. After the calculation of the distances, the reduced data set with the smallest sum value is chosen. It was then observed that the profile of the distances (see Fig. 3.16) is creating a criterion for choosing the SI value associating with a reduced number of input variables and for which the loss of significant quantities of ecological information is minimized. The chosen reduced data set had the higher regression determination coefficient among the examined reduced data sets. The examination if the ecological information is preserved for the new reduced data set was the same as the previous example described by Tison et al. (2005), while the samples of the new data class are used to train the SOM.

**Figure 3.15.** Several classes of the Structuring Index (SI) having different number of variables (species) derived from an original data set of 941 variables (taken by Park et al., 2006).
Figure 3.16. Distance of similarity (chosen to be the Squared Euclidean Distance) for the variables (species) Structuring Index (SI) between the original data set and the new reduced data sets (taken by Park et al., 2006).

3.5. Clustering of the SOM

The division of clusters of the trained SOM prototypes is usually based on the U-matrix (Joo and Jeong, 2005), however two other clustering techniques are also used: the k-means algorithm clustering method and a hierarchical clustering method that is based on dendrogram, which is showing the dissimilarity of the cells in the SOM map (Gevrey et al., 2004). These three clustering methods are graphically represented in the following figure (Figure 3.17). An extensive/in-depth presentation of these three clustering methods applied to the SOM’s prototypes is found in the studies of Vesanto and Alhoniemi (2000); Goncalves et al. (2008).
The first method for clustering of the trained SOM prototypes is based on the U-matrix’s observation, meaning that clusters are detected after visual inspection is performed on the U-matrix (Vesanto and Alhoniemi, 2000). This procedure of clustering is described as in the following figure (Figure 3.18), where the U-matrix of a trained SOM is visualized, and the seven resulted cluster of the measured input data are formed. The dark shades of the grayscale color are indicating large Euclidean distances for the neighboring prototypes, and by that way the borders between the formed clusters are indicated. For example, in the study of Park et al. (2004) the U-matrix algorithm (or else known as Unified matrix algorithm) was applied in order to cluster benthic macroinvertebrates communities in stream ecosystems and to assess the impact of environmental factors on the macroinvertebrates.
Figure 3.18. The U-matrix, where large distances between the neighboring map units are indicated by dark colors of the grayscale. The seven formed (A, B, C, D, E, O₁, O₂) clusters are in circles and the interpolating map units are marked with X. The clusters O₁ and O₂ are corresponding to outliers input data (partially modified from Vesanto and Alhoniemi, (2000)).

The application of the k-means algorithm for clustering the SOM’s prototypes is a relatively easy task, widely used in SOM water quality modelling by many researchers e.g. Garcia and Gonzalez (2004); Astel et al. (2007); Juntunen et al. (2013); Wang et al. (2014); Kim and An (2015); Li et al. (2017). Initially the SOM neural network is trained and then the k-means clustering algorithm -an iterative clustering algorithm that divides a given data set into a number of k clusters-, is applied on the nodes of the trained SOM. The k-means algorithm is partitioning the input data into a specified number of clusters based on the U-matrix (Recknagel et al., 2006b). It must be noted that, when the number of clusters is increased, then the algorithm is sensitive to outliers because the number of samples in each cluster is decreasing (Vesanto and Alhoniemi, 2000).

Meanwhile with the usage of the Davies-Bouldin index it is calculated the optimal number of the k groups (clusters) for the prototypes (nodes) of the trained SOM (Cinar and
The Davies-Bouldin index -for analytical information see the original study of Davies and Bouldin (1979)- is having its minimum value for the optimal number of clusters of the trained SOM (Vesanto and Alhoniemi, 2000; Cinar and Merdun, 2009). The procedure for finding the optimal number of clusters based on the Davies-Bouldin index is graphically illustrated as in the Figure 3.19.

**Figure 3.19.** The Davies-Bouldin clustering index for the k-means clustering algorithm, based on which the optimal number of clusters for the SOM can be calculated by finding for which cluster number corresponds the minimum value of the index (taken by Cinar and Merdun, 2009).

After calculating the Davies-Bouldin clustering index, based on which the optimal number of clusters of the SOM is calculated and symbolized with $n$, then the $n$-cluster structure of the SOM map can be described. For example the SOM clustering that is graphically illustrated as in Figure 3.20, is showing how is done the division/clustering of the SOM map for the water quality data in a $n$-cluster structure, based on the procedure described previously by the Figure 3.19. The study of Cinar and Merdun et al. (2009), which is examining with the use of a SOM model the surface water quality of 4 watersheds in Turkey, in order to investigate the relationships between the water quality variables, is analytically
describing the association of Figure 3.19 and Figure 3.20. In that study the Davies-Bouldin Index was found to be seven and therefore these 4 watersheds were clustered in 7 clusters in the SOM map, allowing the authors to extract information about the variables’ associations in each group, e.g. for group 1 the high water temperature was associated with low rainfall and high Total Dissolved Solids concentrations.

**Figure 3.20.** Clustering of the water quality data set on the SOM map, after calculating the Davies-Bouldin index (taken by Cinar and Merdun, 2009).

In similar SOM modelling studies (e.g. Zelazny et al., 2011; An et al., 2016; Li et al., 2017; Hadjisolomou et al., 2018) with the one mentioned above, which are examining the surface water quality and the k-means algorithm is applied on the trained SOM prototypes and also the Davies-Bouldin Index is calculated, in order to find the optimal clustering that corresponds for the under study watersheds; an additional step is proceed. For each resulted cluster the statistical evaluation is performed, in order to find the statistical properties (mean, maximum, minimum, standard deviation) that are corresponding for each cluster and by that way to extract useful conclusions about the parameters’ behavior. As stated by Hadjisolomou
et al. (2018) the role of water temperature and the seasonality effect was having a crucial role on the shallow eutrophic lake Mikri Prespa, while the deep meso-eutrophic lake Megali Prespa was less prone to the seasonality effect.

The clustering of the SOM’s prototypes can be also done with the use of the hierarchical clustering method that is based on a dendrogram and the usage of the Ward linkage method based on the Euclidean distance (Chon, 2011). Based on the Ward’s linkage as a clustering method the weight vectors for each neuron of the output layer are used (Choi et al., 2014). Several SOM-based eco-hydrological modelling studies are using this hierarchical clustering method like the ones of Tison et al. (2005); Oh et al. (2007); Lek-Ang et al. (2007); Astel et al. (2007); Choi et al. (2014); Kwon et al. (2012), Hadjisalomou et al. (2018). The graphical illustration of the use of hierarchical cluster analysis on the trained SOM’s prototypes is given as in the Figure 3.21(taken by the study of Sroczynska et al. (2017)). It must be noted that the SOM model of the study of Sroczynska et al. (2017) - that is examining the macroinvertebrates species in the Quarteira stream system (Portugal)- managed to investigate the variety of traits and preferences (meaning biotic data and abiotic conditions) of macroinvertebrates, while the macroinvertebrates community composition was varying throughout the large monitored study area.

The clustering of the SOM prototypes based on the different clustering methodologies can give useful results and reveal several mechanisms related with the function of the aquatic ecosystem. However as stated by Vesanto and Alhoniemi (2000) the clusters examination shouldn’t be limited only to examine their statistical properties (e.g. mean, medium, maximum, minimum), but also other aspects of the clustering results should be taken into consideration like: if the is cluster an outlier or does it have sub-clusters; which factors are making the data that belong to that cluster to be different from the rest data and the neighboring clusters.
Figure 3.21. The SOM map for the assigned macroinvertebrates samples (left figure) and the associated dendrogram resulted after applying hierarchical cluster analysis on the SOM’s nodes (right figure). Two main clusters are distinguished (X, Y) and five sub-clusters (X₁, X₂, Y₁, Y₂, Y₃) with the use of hierarchical cluster analysis. The indexing A₁-F₄ is symbolizing the neurons order/arrangement on the SOM map; the RA, RM; RQ; FS, FB, BG are symbolizing the names of the different river sampled and the Arabic numbers stand for the day and month (taken by Sroczynska et al., 2017).

3.6. Application of the SOM neural network for prediction

The SOM neural network is mainly used as a data clustering methodology and only few researchers have applied the SOM for prediction, in contrast with the supervised ANNs that are mainly used for prediction. The SOM is flexible to predict time series and because of its visualization abilities allow the user to understand the dynamics of the underlying process.
in a better way than the MLP neural networks (Barreto, 2007). According to Simon et al. (2003) the process for selecting the best topology of the SOM model when it is used for time series prediction is not really simple in practice, therefore they have proposed a methodology for selecting the best SOM neural network topology based on a normalized error (NSSE) criterion as given by the equation below (Eq. 3.16):

$$NSSE_{n1n2} = \frac{\sum (x(t+1) - \hat{x}(t+1))^2}{\sum (x(t+1) - E[x(t+1)])^2}$$  \hspace{1cm} (Eq. 3.16)

where $x(t+1)$ the real value; $\hat{x}(t+1)$ the predicted value; $E[.]$ symbolizes the mean and is estimated on the measured values of the time series; the $n1n2$ index is associated with the prototypes number and the $n1$ and $n2$ are predefined numbers.

Especially for the thematic field of eco-hydrological modelling only a relatively small number of studies using the SOM neural network for time prediction exist, like the ones of Lee and Scholz (2006b); Zhang et al. (2008); Zhang et al. (2009). The SOM neural network for times series prediction has great significance for hydrological modelling as derived from the literature. The SOM can simulate parameters that are not time-efficient, using as model's input parameters more time-efficient variables.

For example, in the study of Zhang et al. (2009) the BOD outflow concentrations of an integrated constructed wetland of an agroecosystem treating agricultural runoff were simulated with the use of a SOM model. As stated by the above authors the time consuming to measure BOD outflow concentrations were predicted by inflow variables that were quicker and easier to measure. The constructed SOM model managed to simulate well and with good accuracy the BOD outflow concentrations (see Fig. 3.22) for the test set, while the correlation coefficient ($R$) was high ($R=0.96$) with small $MRAE$ and $MSE$ during the training. Therefore, based on the SOM neural network performance evaluation it was concluded that the measured
BOD outflow parameter that is time-consuming (at least five days) could be replaced with the simulated BOD outflow parameter.

Figure 3.22. Time series prediction for the measured and the predicted biological oxygen demand (BOD) outflow data for the test set (taken by Zhang et al., 2009).

The SOM neural network that can be used for time series prediction can be also successfully applied in the field ecohydrological sciences for the prediction of variables that are not cost-efficient (expensive to measure). For example, in their study Zhang et al. (2008) simulated the nutrient removal performance in integrated constructed wetlands by applying a SOM neural network. The SOM model managed to predict with good accuracy the ammonia-nitrogen and SRP outflow concentrations with the use of variables (WT; EC; DO) that can be measured less cost-effective.

The SOM Toolbox for Matlab developed by the Laboratory of Information and Computer Science at Helsinki University of Technology (Vesanto et al., 1999) is the main tool that is used in the majority of SOM modelling studies in the field of eco-hydrology e.g. Lee
and Scholz (2006a); Recknagel et al. (2006a); Recknagel et al. (2006b); Jeong et al. (2010); Choi et al. (2014); Olkowska et al. (2014); Sroczynska et al. (2017); Hadjisalomou et al. (2018). The SOM toolbox for Matlab is freely available online at http://www.cis.hut.fi/projects/somtoolbox.

In conclusion, the SOM neural network has been proved to be an efficient way to evaluate the complex relationships between water quality variables. The SOM model is characterized as an excellent practical tool, which is good in prediction and that is not affected by possible computational barriers like missing values and incomplete data sets (Zhang et al., 2009). Therefore, it had started to be used as a management tool in order to solve several water quality problems like agricultural, industrial and municipal water pollution (Cinar and Merdun, 2009). Besides SOM’s ability to model complex non-linear relationships more reliable than other modelling methodologies, the advanced visualization abilities of the SOM neural network are making to be a superior modelling tool when comparing with other classical modelling techniques (Astel et al., 2007; Hadjisalomou et al., 2018).

3.7. References

Al-Mudhaf, H. F.; Astel, M. A.; Selim, M. I.; Abu-Shady, A.-S. Self-organizing map approach in assessment spatiotemporal variations of trihalomethanes in desalinated drinking water in Kuwait. *Desalination* 2010, 252, 97-105.

An, Y.; Zou, Z.; Li, R. Descriptive Characteristics of Surface Water Quality in Hong Kong by a Self-Organising Map *Int. J. Environ. Res. Public Health* 2016, 13, 115; doi:10.3390/ijerph13010115.
Astel, A.; Tsakovski, S.; Barbieri, P.; Simeonov, V. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research* 2007, 41, 4566-4578.

Barreto, G. Time Series Prediction with the Self-Organizing Map: A Review. In *Perspectives of Neural-Symbolic Integration*; Hammer, B.; Hitzler, P., Eds; Springer-Verlag: Berlin Heidelberg, Germany, 2007, pp 135-158.

Barreto, M.A.; Perez-Uribe, A. Improving the Correlation Hunting in a Large Quantity of SOM Component Planes. In *Lecture Notes in Computer Science: Artificial Neural-Networks–ICANN 2007*; de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D., Eds.; Springer: Berlin, Germany, 2007; Volume 4669, pp. 379–388.

Bieroza, M.; Baker, A.; Bridgeman, J. Exploratory Analysis of excitation-emission matrix fluorescence spectra with self-organizing maps as a basis for determination of organic matter removal efficiency at water treatment works. *Journal of Geophysical Research* 2009, 114, G00F07, doi: 10.1029/2009JG000940

Buscema, M. How to perform data mining. In *Applications of Mathematics in Models, Artificial Neural Networks and Arts*; Capecchi, V.; Buscema, M.; Contucci, P.; D’Amore, B., Eds; Springer-Verlag: Berlin, Germany, 2010.

Cereghino, R.; Park, Y.-S. Review of the Self-Organizing Map (SOM) approach in water resources. *Environmental Modelling & Software* 2009, 24, 945-947.

Choi, B.-Y.; Yun, S.-T.; Kim, K.-H.; Kim, J.-W.; Kim, H.M.; Koh, Y.-K. Hydrogeochemical interpretation of South Korea groundwater monitoring data using Self-Organizing Maps. *Journal of Geochemical Exploration* 2014, 137, 73-84.

Chon, T.-S. Self-Organizing Maps applied to ecological sciences. *Ecological Informatics* 2011, 6, 50-61.
Cho, H.-S.; Choi, K.-H.; Lee; S.-D.; Park, Y.-S. Characterizing habitat preference of Eurasian river otter (*Lutra lutra*) in streams using a self-organizing map. *Limnology 2009*, *10*, 203-213.

Cinar, O.; Merdun, H. Application of an unsupervised artificial neural network technique to multivariant surface water quality data. *Ecol Res 2009*, *24*, 163-173.

Davies, D.L.; Bouldin, D.W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence 1979*, *PAMI-1*, 224-227.

Dong, Y.; Scholz, M., Harrington, R. Statistical Modeling of Contaminants Removal in Matura Integrated Constructed Wetlands Sediments. *Journal of Environmental Engineering 2012*, *138*, 1009-1017.

Farsadnia, F.; Rostami Kamrood, M.; Moghaddam Nia, A.; Modarres, R.; Bray, M.T et al. Identification of homogenous regions for regionalization of watersheds by two-level self-organizing feature maps. *Journal of Hydrology 2014*, *509*, 387-397.

Garcia, H.L.; Gonzalez, I.M. Self-organizing map and clustering for wastewater water treatment monitoring. *Engineering Applications of Artificial Intelligence 2004*, *17*, 215-225.

Giraudel, J.L., Lek, S. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling 2001*, *146*, 329-339.

Gevrey, M.; Rimet, F.; Park, Y.-S.; Giraudel, J.-L.; Ector, L.; Lek, S. Water quality assessment using diatom assemblages and advanced modelling techniques. *Freshwater Biology 2004*, *49*, 208-220.

Goncalves, M.L.; Netto, M.L.A.; Costa, J.A.F.; Zullo Junior, J. An unsupervised method for classifying remotely sensed images using Kohonen self-organizing maps and agglomerative hierarchical clustering methods.
Hadjisolomou, E.; Stefanidis, K.; Papanastasiou, G. Papastergiadou, E. Assessment of the Eutrophication-Related Environmental Parameters in Two Mediterranean Lakes by Integrating Statistical Techniques and Self-Organizing Maps. *International Journal of Environmental Research and Public Health* 2018, 15(3):547, doi: 10.3390/ijerph15030547.

Jahan, R.; Choi, H. C.; Park, Y. S.; Park, Y. C.; Seo, J. H.; Choi, J. K. Implementation of the Self-organizing Maps (SOM) to analyses of environmental parameters and phytoplankton biomass in a microtidal estuary and artificial lake. *Journal of the Marine Biological Association of the United Kingdom* 2013, 93, 1-12.

Jeong, K.-S.; Hong, D.-G.; Byeon, M.-S.; Jeong, J.-C.; Kim, H.-G. et al. Stream modification patterns in a river basin: Field survey and self-organizing map (SOM) application. *Ecological Informatics* 2010, 5, 293-303.

Joo, G.J.; Jeong, K.-S. Modelling community changes of cyanobacteria in a flow regulated river (the lower Nakdong River, S. Korea) by means of a Self-Organizing Map (SOM). In *Modelling Community Structure in Freshwater Ecosystems*; Lek, S.; Scardi, M.; Verdonschot, P.F.M.; Descy, J.-P.; Park, Y.-S. Eds; Springer-Verlag: Berlin Heidelberg, Germany, 2005.

Juntunen, P.; Liukkonen, M.; Lehtola, M.; Hiltunen, Y. Cluster analysis by self-organizing maps: An application to the modeling of water quality in a treatment process. *Applied Soft Computing* 2013, 13, 3191-3196.

Kangas, J.; Kohonen, T. Developments and applications of the self-organizing map and related algorithms. *Mathematics and Computers in Simulation* 1996, 41, 3-12.

Kangur, K.; Park, Y.-S.; Kangur, A.; Kangur, P.; Lek, S. Patterning long-term changes of fish community in large shallow Lake Peipsi. *Ecological Modelling* 2007, 203, 34-44.
Kim, J.Y.; An, K.-G. Integrated Ecological River Health Assessments, Based on Water Chemistry, Physical Habitat Quality and Biological Integrity. *Water* 2015, 7, 6378-6403; doi:10.3390/w7116378.

Kohonen, T. Analysis of a simple self-organising process. *Biol Cybern* 1982, 44, 135-140.

Kohonen, T. *Self-organising maps*; Springer-Verlag: Berlin Heidelberg, Germany, 2001.

Kohonen, T. The self-organising map. *Neurocomputing* 1998, 21, 1-6.

Kwon, Y.-S.; Li, F.; Chung, N.; Bae, M.-J.; Hwang, S.-J. et al. Response of Fish Communities to Various Environmental Variables across Multiple Spatial Scales. *Int. J. Environ. Res. Public Health* 2012, 9, 3629-3653; doi:10.3390/ijerph9103629

Lee, B.-H., Scholz, M. A comparative study: prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks. *Water, Air, and Soil Pollution* 2006b, 174, 279-301.

Lee, B.-H.; Scholz, M. Application of the self-organizing map (SOM) to assess the heavy metal removal performance in experimental constructed wetlands. *Water Research* 2006a, 40, 3367-3374.

Lek-Ang, S.; Park, Y.-S.; Ait-Mouloud, S.; Deharveng, L. Collembolan communities in a peat bog versus surrounding forest analyzed by using self-organizing map. *Ecological Modelling* 2007, 203, 9-17.

Li, X.; Sha, Z.; Wang, Z.-L. Chlorophyll-A Prediction of Lakes with Different Water Quality Patterns in China Based on Hybrid Neural Networks. *Water* 2017, 9(7), 524; doi:10.3390/w9070524

Liu, Y.; Weisberg, R.; Mooers, C. Performance evaluation of the self-organizing map for feature extraction. *Journal of Geophysical Research* 2006, 111, doi:10.1029/2005JC003117
Liu, Y.; Weisberg, R. A review of self-organizing map applications in meteorology and oceanography. In *Self-Organizing Map Applications and Novel Algorithm Design*; Mwasiagi, J.I. Eds; InTech: Rijeka, Croatia, 2011.

Natita, W.; Wiboonsak, W.; Dusadee, S. Appropriate Learning Rate and Neighborhood Function of Self-organizing Map (SOM) for Specific Humidity Pattern Classification over Southern Thailand. *International Journal of Modeling and Optimization* 2016, 6, 61-65.

Nourani, V.; Baghanam, A.H.; Adamowski, J.; Gebremichael, M. Using self-organizing maps and wavelet transforms for space–time pre-processing of satellite precipitation and runoff data in neural network based rainfall–runoff modeling. *Journal of Hydrology* 2013, 476, 228-243.

Oh, H.-M.; Ahn, C.-Y.; Lee, J.-W.; Chon, T.-S.; Choi, K.H.; Park, Y.-S. Community patterning and identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using artificial neural networks. *Ecological Modelling* 2007, 203, 109-118.

Olkowska, E.; Kudlak, B.; Tsakovski, S.; Ruman, M.; Simeonov, V.; Polkowska, Z. Assessment of the water quality of the Klodnica River catchment using self-organizing maps. *Science of the Total Environment* 2014, 476-477, 477-484.

Park, Y. -S.; Chon, T.-S.; Kwak, I.-S.; Lek, S. Hierarchical community classification and assessment of aquatic ecosystems using artificial neural networks. *Science of the Total Environment* 2004, 327, 105-122.

Park, Y. -S.; Gevrey, M.; Lek, S.; Giraudel, J. L. Evaluation of relevant species in communities: development of structuring indices for the classification of communities using a self-organizing map. In *Modelling Community Structure in Freshwater*
Peeters, L.; Bacao, F.; Lobo, V.; Dassargues, A. Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen's Self-Organizing Map. *Hydrol. Earth Syst. Sci.* 2007, 11, 1309-1321.

Plociennik, M.; Kruk, A.; Michczynska, D.; Birks, J. Kohonen Artificial Neural Networks and the IndVal Index as supplementary tools for the quantitative analysis of paleoecological data. *Geochronometria* 2015, 42, 189-201.

Recknagel, F.; Talib, A.; van der Molen, D. Phytoplankton community dynamics of two adjacent Dutch lakes in response to seasons and eutrophication control unravelled by non-supervised artificial neural networks. *Ecological Informatics* 2006, 1, 277-285.

Recknagel, F.; Cao, H.; Kim, B.; Takamura, N.; Welk, A. Unravelling and forecasting algal population dynamics in two lakes different in morphometry and eutrophication ny neural and evolutionary computation. *Ecological Informatics* 2006, 1, 133-151.

Scholz, M. Modeling Complex Wetland Systems. In *Wetland Systems*; Springer-Verlag: Berlin, Germany, 2011. pp 217-231.

Simon, G.; Lendasse, A.; Cottrell, M.; Verleysen, M. Long-Term Time Series Forecasting Using Self-Organizing Maps: The Double Vector Quantization Method. In ANNPR’2003 proceedings: Florence, Italy, 2003.

Sroczynska, K.; Claro, M.; Kruk, A.; Wojtal-Frankiewicz, A.; Range, P.; Chicharo, L. Indicator macroinvertebrate species in a temporary Mediterranean river: Recognition
of patterns in binary assemblage data with a Kohonen artificial neural network. 

*Ecological Indicators* 2017, 73, 319-330.

Stefanovic, P; Kurasova, O. Visual Analysis of Self-organizing maps. *Nonlinear Analysis: Modelling and Control* 2011, 16, 488–504.

Tison, J.; Park, Y. -S.; Coste, M.; Wasson, J. G.; Ector, L. et al. Typology of diatom communities and the influence of hydro-ecoregions: A study on the French hydrosystem scale. *Water Research* 2005, 39, 3177-3188.

Tota-Maharaj, K.; Scholz, M. Modelling of permeable pavements for treatment of urban runoff using self-organizing maps. *Environmental Engineering and management Journal* 2013, 12, 2273-2287.

Tsai, W. -P.; Huang, S. -P.; Cheng, S. -T.; Shao, K.-T.; Chang, F.-J. A data mining framework for exploring the multi-relation between fish species and water quality through the self-organizing map. *Science of the Total Environment* 2017, 579, 474-483.

Vesanto, J.; Ahola, J.; Hunting for correlations in Data Using the Self-Organizing Map. Proceeding of the International ICSC Congress on Computational Intelligence Methods and Applications, pp. 279-285, 1999.

Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. Self-organizing map in Matlab: the SOM toolbox. In: Proceedings of the Matlab DSP Conference, November 1999, Espoo, Finland, pp. 35-40.

Vesanto, J. SOM-based data visualization methods. *Intelligent Data Analysis* 1999, 3, 111-126.

Vesanto, J.; Himberg, J.; Alhoniemi, E.; Parhankangas, J. SOM toolbox for matlab 5 technical report A57. Helsinki, Finland: Neural Networks Research Centre, Helsinki University of Technology; 2000.
Vesanto, J.; Alhoniemi, E. *IEEE Transactions on Neural Networks* 2000, *11*, 586-600.

Voutilainen, A.; Rahkola-Sorsa, M.; Parviainen, J.; Huttunen, M.J.; Viljanen, M. Analysing a large dataset on long-term monitoring of water quality and plankton with the SOM clustering. *Knowledge and Management of Aquatic Ecosystems* 2012, *406*, 04-16.

Ramadas, M.; Ostermann, S.; Tjaden, B. Detecting anomalous Network Traffic with Self-Organizing Maps. In *Recent Advances in Intrusion Detection* Vigna, G.; Jonsson, E.; Kruegel, C., Eds; Springer-Verlag: Berlin, Germany, 2003.

Wang, B.; Li, H.; Sun, D. Social-Ecological Patterns of Soil Heavy Metals Based on a Self-Organizing Map (SOM): A Case Study in Beijing, China *Int. J. Environ. Res. Public Health* 2014, *11*, 3618-3638; doi:10.3390/ijerph110403618.

Zelazny, M.; Astel, A.; Wolanin, A.; Malek, S. Spatiotemporal dynamics of spring and stream water chemistry in a high-mountain area. *Environmental Pollution* 2011, *159*, 1048-1057.

Zhang, L.; Scholz, M.; Mustafa, A.; Harrington, R. Assessment of the nutrient removal performance in integrated constructed wetlands with the self-organizing map. *Water Resources* 2008, *42*, 3519–3527.

Zhang, L.; Scholz, M.; Mustafa, A.; Harrington, R. Application of the self-organizing map as a prediction tool for an integrated constructed wetland agroecosystem treating agricultural runoff. *Bioresource Technology* 2009, *100*, 559-565.

Zhang, Y.; Liang, X.; Wang, Z.; Xu, L. A novel approach combining self-organizing map and parallel factor analysis for monitoring water quality of watersheds under non-point source pollution. *Scientific Reports* 2015, *5*:16079/ DOI: 10.1038/srep16079.
4. Artificial Neural Networks and other popular modelling techniques.

Among the most popular methods for modelling water quality parameters are the Multiple Linear Regression (MLR), Principal Component Analysis (PCA) and Factor Analysis. Some of their limnological applications and their brief description are sited below. Also, some basics of statistical theory necessary for the needs of many modelling studies are given. The way that these methodologies are combined with ANNs in order to resolve some modelling issues is also examined. Finally, ANN models are compared against these broadly used modelling methodologies applied for water quality modelling, in order to evaluate/compare them as modelling tools.

4.1. Applications of Statistical Techniques in water quality modelling

Statistical techniques are widely applied in water quality parameters modelling. Many modelling studies exist, which are examining the impact of the environmental parameters on lakes and the methods of PCA and MLR are applied for this purpose. The modelling results found by these kinds of studies are very useful providing a good evaluation tool in order to assess the limnological parameters and some of their interactions. For example, Papatheodorou et al. (2002) used PCA to examine the heavy metals concentrations in sediments in a shallow lagoon and identified the main pollution sources. In another study of Stefanidis & Papastergiadou (2012) PCA was applied in order to investigate the relationships between water quality parameters, aquatic macrophytes and the morphological characteristics of nineteen selected lakes of Greece. The results of PCA revealed that the lake volume and the altitude were the key variables based on which the lakes were discriminated. The study of
Kagalou and Leonardos (2009) used PCA to group the main Greek lakes according to their morphological characteristics and to reveal correlations between the lakes’ variables.

The MLR models are also widely used in the field of water quality research, where they are managing to represent with success difficult modelling tasks. For that reason, several studies were carried out using MLR examining water quality related problems. Such study is the one carried out by Chu et al. (2013), in which the MLR model was applied in order to examine the relationship between the water quality parameters and the land cover changes after typhoon events in a watershed obtained by satellite images. In the study of Zhang et al. (2012) MLR methods are used for associating the impact of agricultural land use intensity on the water quality of five Beijing's mountainous watersheds.

4.2. Correlation Coefficient

The Pearson correlation symbolized as correlation coefficient ($r$) is the most widely used measure to associate two variables for continues types data and provides a measure of the linear association of two variables (Marques de Sa, 2007) and is calculated based on the type:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

(Eq. 4.1)

where $\bar{x}$ is the mean of variable $x$; $\bar{y}$ is the mean of variable $y$; $n$ the number of measured samples. When the correlation coefficient equals +1 then it is indicated that there is a perfect sympathetic linear relationship between the two variables, while when the correlation coefficient equals -1 the there is a linear antipathetic relationship between the two variables.
(Marques de Sa, 2007; Reimann, 2008). Prior calculating Pearson correlation coefficient, it is recommended that the data to be transformed, because the Pearson method is very sensitive to data outliers and deviations from the main structure of the data (Reimann, 2008).

Another widely used type of correlation coefficient is the Spearman's correlation coefficient and is symbolized as $r_s$. The formula that gives the Spearman's correlation coefficient is having the follow form (Mukaka, 2012):

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)} \quad \text{(Eq. 4.2)}$$

where $d_i^2$ is the difference in ranks for variable $x$ and variable $y$; $n$ is the observations number.

The alternative form of the Spearman's correlation coefficient is as bellow (Millard & Neerchal, 2000):

$$r_s = \frac{\sum_{i=1}^{n} (R_{x_i} - \bar{R}_x)(R_{y_i} - \bar{R}_y)}{\sqrt{\sum_{i=1}^{n} (R_{x_i} - \bar{R}_x)^2 \sum_{i=1}^{n} (R_{y_i} - \bar{R}_y)^2}} \quad \text{(Eq. 4.3)}$$

where the $R_x$ is the rank for variable $x$ and $R_y$ rank for variable $y$; $n$ is the observations number.

In contrast with the Pearson correlation coefficient the Spearman’s correlation coefficient measures any kind of monotonic relationship between $x$ and $y$ and is not sensitive to the outliers (Mukaka, 2012; Millard & Neerchal, 2000).
4.3. Principal Components Analysis

Principal Components Analysis (PCA) is a multivariate statistical technique that is used to reduce the number of variables of a data set to a smaller number of variables without information loss of the initial data set (Hou et al., 2015). A new set of variables called principal components are obtained from applying PCA on the data set. New uncorrelated (orthogonal) variables are created after performing VARIMAX rotation, which they are linear combination of the initial variables (Chen et al., 2016) and eigenvalues and eigenvectors are extracted from the covariance matrix of the initial data set. Prior applying PCA it is recommended the data to be standardized in order to be reflected in a uniform measurement system across all variables (Hou et al., 2015). The resulted principal components are represented as below (Duan et al., 2015):

\[ F_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p \]
\[ F_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p \]
\[ \ldots \]
\[ \ldots \]
\[ F_p = a_{p1}x_1 + a_{p2}x_2 + \ldots + a_{pp}x_p \]  
\hspace{3cm} (Eq. 4.4)

or represented with their general form that is:

\[ F_j = a_{ji}x_1 + a_{j2}x_2 + \ldots + a_{jp}x_p \]  
\hspace{3cm} (Eq. 4.5)
where $F_j$ symbolizes the component score; $x_j$ is the $j$th measured value of a variable; $p$ is the total number of variables; $a_{jp}$ is the loading of the $p$th variable in the $j$th factor; $j=1, 2, \ldots, p$.

Some rules suggesting the number of how many principal components to retain are given as below (Jakson, 1993; Rencher, 2002; Marques de Sa, 2007):

1. Select the principal components that explain a certain percentage of the total variance (say, 80%).
2. The Guttman-Kaiser criterion retains these principal components that their eigenvalues are greater than the average of the eigenvalues (over 1 for standardised data).
3. The scree test uses the scree-plot (graph of the eigenvalues) and search for a break between the large and small eigenvalues.
4. The broken-stick model, which assumes that the expected distribution of the eigenvalues will follow a broken stick distribution if the total variance is divided randomly amongst the various components. The observed eigenvalues are discarded if they do not exceed the eigenvalues found by the broken-stick model. The eigenvalues generated by the broken-stick model are calculated based on the equation:

$$b_k = \sum_{i=k}^{p} \frac{1}{i} \quad \text{(Eq. 4.6)}$$

where the $p$ is the number of variables; $b_k$ is the size of the eigenvalue for the $k$th principal component.

More specific for Rule 1 it is proposed as the percentage of the total variance explained by the principal components should account approximately 75% (Duan et al., 2015). The proportion
of variance explained by the first $k$ components is calculated as follow (Hou et al., 2015; Rencher, 2002):

\[
Proportion\ of\ variance = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad (Eq.\ 4.7)
\]

where $\lambda_i$ is the eigenvalue of the $i$th principal component.

### 4.4. Factor Analysis

Factor Analysis is multivariate statistical method used to find a small number of factors from a data set of many correlated variables and the dimensionality of the initial data set can be reduced with the application of factor analysis (Astel et al., 2004). Factor Analysis differs from PCA because Factor Analysis imposes a strict structure of a fixed number of common factors, while PCA determines $p$ factors in a decreasing order of importance (Hardle and Simar, 2003). Also, a major difference between PCA and Factor Analysis is based on the fact that in the case of PCA the principal components are are linear combinations of the original variables, while in the case of Factor Analysis the original variables are linear compilations of the factors (Rencher, 2002).

Before contacting Factor Analysis, it is advised to exam if the data are suitable for contacting Factor Analysis, with the use the correlation matrix and to check if most of the correlations are over 0.3, and also with the use of some tests. The Bartlett's test of sphericity that examines the null hypothesis that the variables are uncorrelated must be met and the Kaiser-Meyer-Olkin measure must be greater than 0.6 (Foster et al., 2006). The Factor Analysis model represents each variable as a linear combination of underlying common
factors \( f_1, f_2, \ldots, f_m \) with a unique accompanying error term for each variable and the Factor Analysis model is given as below (Rencher, 2002):

\[
y_1 = \lambda_{11} f_1 + \lambda_{12} f_2 + \ldots + \lambda_{1m} f_m + \epsilon_1 \\
y_2 = \lambda_{21} f_1 + \lambda_{22} f_2 + \ldots + \lambda_{2m} f_m + \epsilon_2 \\
\vdots \\
y_p = \lambda_{p1} f_1 + \lambda_{p2} f_2 + \ldots + \lambda_{pm} f_m + \epsilon_p
\]

(Eq. 4.8)

where the \( \epsilon_i \) is the error term; the coefficients \( \lambda_{ij} \) are called loadings and served as weights, showing how each variable depends from the loading.

Generally, the number of extracted factors must be less than the number of measured variables (Astel et al., 2004). The three most used selection methods for the number of factors to be extracted are very similar to those used for PCA for choosing the number of principal components to retain and they are described by Brown (2006) as follow:

1. **The Kaiser-Guttman rule**: also known as the Kaiser criterion or the eigenvalues greater than one rule, which uses eigenvalues derived from the input correlation matrix.

2. **The scree test**: is based on the inspection of a graph of the eigenvalues (x axis) and the factors (y axis) and to determine where the graph changes slope.

3. **Parallel Analysis**: is based on a scree plot of the eigenvalues derived from the sample data against eigenvalues derived from factoring a random number data set.
After the factors number is decided, the extracted factors are rotated, aiming with rotation to obtain factors with high correlation for some variables and no correlation with the others (Basilevsky, 1994), and providing by that way a more informative interpretation of the factors (Hardle and Simar, 2003). Several rotation methods exist like the Varimax, which is an orthogonal rotation and results in orthogonal rotated factors and the Promax; Oblimin; Quartimim methods which are oblique rotation methods, meaning the rotated factors aren't orthogonal (Reimann et al., 2008).

The process of Factor Analysis can be summarized into the following steps described by Foster et al. (2006):

1. Examine if Factor Analysis can be applied.
2. Decide how many factors there are.
3. Carry out Factor Analysis using the number of factors found previously.
4. Examine the results of Factor Analysis and evaluate the outcome how clear it is.
5. Repeat the Factor Analysis using an orthogonal rotation of the factors and examine if the outcome is clearer of that produced in step 3.
6. Repeat the Factor Analysis using an oblique rotation of the factors and examine if the outcome is clearer of that produced in step 5.
7. Compare the results of Factor Analysis with the oblique and orthogonal rotations and find the one with the simpler structure.
8. Interpret the factors with the results of the simpler structure found in step 7.
4.5. Cluster analysis

Cluster Analysis is a multivariate statistical method that is classifying parameters into clusters (groups) based on their similarity and this classification is done without any prior assumptions regarding the structure or underlying patterns of the original data set (Kura et al., 2013). Cluster Analysis is referred as classification, pattern recognition (unsupervised learning) and numerical taxonomy (Rencher, 2002). Objects are clustered based on their similarity. An index of similarity between each pair of observations is used by many clustering methods. The degree of similarity among objects is measured with the similarity (proximity) or distance matrices and high levels of similarity among objects are having large value in the similarity matrix and small value in the distance matrix (Gore, 2000). The dissimilarity measure is usually calculated based on the squared Euclidean distance, although some other dissimilarity measures exist like the city-block distance (or Manhattan metric); Chebychev distance; Angular distance; Canberra or Mahalonobis distance (Sarstedt & Mooi, 2004).

Cluster methods are divided into four categories: hierarchical methods; partitioning methods; overlapping cluster procedures; ordination techniques (Hardle and Simar, 2003; Gore, 2000). The two most widely used types of clustering methods are the hierarchical (or nested) and the partitioning (or unnested) methods. In contrast to hierarchical algorithms, the partitioning methods have the number of their clusters pre-determined, meaning known a priori (Reimann et al., 2008). The main difference between hierarchical clustering and portioning clustering is that with the hierarchical clustering techniques once the objects are assigned to the groups this assignment cannot be changed, while with the partitioning techniques the assignment of objects may change during the algorithm application (Hardle and Simar, 2003).

The most famous partitioning algorithm is the K-means algorithm. The main idea of the k-means algorithm is to minimise the average of the squared distances between the
observations and their cluster centres (centroids). This process is starting with the algorithm having assigned the observations to their closest centroid, recomputes the cluster centroids and iteratively reallocates the data points to the closest centroid (Reimann et al., 2008). The K-means algorithm is using the within-clusters variation as a measure to form homogenous clusters, in such way that the within-cluster variation to be minimized (Tan et al., 2006). The steps being followed by the K-means algorithm procedure are synopsized as in Table 4.1.

Table 4.1. Steps of the Basic K-means algorithm (taken by Tan et al., 2006)

| Basic K-means algorithm |
|------------------------|
| 1: Select $K$ points as initial centroids. |
| 2: repeat |
| 3: Form $K$ clusters by assigning each point to its closest centroid. |
| 4: Recompute the centroid of each cluster. |
| 5: until centroids do not change. |

The hierarchical clustering methods are divided into the well-known agglomerative methods and divisive (or splitting) methods. At the beginning of the clustering process the hierarchical agglomerative methods treat every object as a unique cluster, while the hierarchical divisive methods as a single cluster and proceed by establishing smaller clusters (Gore, 2000). The agglomerative hierarchical methods begin with $n$ items and end with a single cluster containing all $n$ items. The divisive hierarchical method starts with a single cluster containing $n$ objects and by dividing the resulting subgroups at the end $n$ clusters are formed with one object each (Rencher, 2002) as represented in the figure (Fig. 4.1).
The Agglomerative Hierarchical clustering techniques start from an object that forms a cluster and compute the distance matrix for the clusters and then join the clusters that have the smallest distance, with this step to be repeated until all objects are united in one cluster (Hardle and Simar, 2003). The steps being followed by the Agglomerative Hierarchical clustering methods are synopsized below in Table 4.2.

Several different methods exist to link two clusters during the procedure being followed by the agglomerative hierarchical algorithm (Reimann et al., 2008). These different agglomerative hierarchical procedures are measuring the distance from a newly formed cluster to a certain object belonging to another cluster with a different way. The most popular of these agglomerative hierarchical procedures are described as below (Rencher, 2002; Tan et al., 2006):
Table 4.2. Steps of the Basic Agglomerative Hierarchical clustering algorithm (taken by Tan et al., 2006)

**Basic Agglomerative Hierarchical clustering algorithm**

1: Compute the proximity matrix, if necessary

2: repeat

3: Merge the closest two clusters.

4: Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.

5: until only one cluster remains.

- **Single linkage** (nearest neighbor): The minimum distance between any two objects in the two clusters is the distance between the two clusters (see Equation 4.9).

\[
D(A, B) = \min(d(y_i, y_j)) \quad (\text{Eq. 4.9})
\]

where \(d(y_i, y_j)\) is usually the Euclidean distance; \(y_i\) object of cluster \(A\); \(y_j\) object of cluster \(B\).

- **Complete linkage** (furthest neighbor): The maximum distance between any two objects in the two clusters is the the distance between the two clusters (see Equation 4.10).

\[
D(A, B) = \max(d(y_i, y_j)) \quad (\text{Eq. 4.10})
\]

where \(d(y_i, y_j)\) is usually the Euclidean distance; \(y_i\) object of cluster \(A\); \(y_j\) object of cluster \(B\).
• **Average linkage**: The average distance between all pairs of the two clusters' objects is the distance between the two clusters (see Equation 4.11).

\[
D(A, B) = \frac{1}{n_An_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j) \quad (Eq. 4.11)
\]

where \( d(y_i, y_j) \) is usually the Euclidean distance; \( y_i \) object of cluster \( A \); \( y_j \) object of cluster \( B \); \( n_A n_B \) the distances between the \( n_A \) points in cluster \( A \) and the \( n_B \) points in cluster \( B \).

• **Centroid**: The distance between the two clusters is the distance between the centroids (mean vectors) of the two clusters (see Equation 4.12).

\[
D(A, B) = d(\overline{y}_A, \overline{y}_B) \quad (Eq. 4.12)
\]

where

\[
\overline{y}_A = \sum_{i=1}^{n_A} \frac{y_i}{n_A} \quad (Eq. 4.13)
\]

and the \( \overline{y}_A \) is the mean vector for the observation vectors in \( A \); \( \overline{y}_B \) is the mean vector for the observation vectors in \( B \); the \( n_A \) points in cluster \( A \).

• **Ward** (incremental sum of of squares method): The Ward method is using the within-cluster (squared) distances and the between-cluster (squared) distances. The aim of the Ward method is to unify clusters in a way that the resulting clusters are homogenous as possible. The procedure being followed by the Ward clustering algorithm is to merge the two clusters \( A \) and \( B \) and to minimize the increase in \( SSE \), symbolized as \( I_{AB} \) and defined based on the following equations:
\[ I_{AB} = SSE_{AB} - (SSE_A + SSE_B) \]  
(Eq. 4.14)

\[ SSE_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)(y_i - \bar{y}_A) \]  
(Eq. 4.15)

\[ SSE_B = \sum_{i=1}^{n_B} (y_i - \bar{y}_B)(y_i - \bar{y}_B) \]  
(Eq. 4.16)

\[ SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})(y_i - \bar{y}_{AB}) \]  
(Eq. 4.17)

where

\[ \bar{y}_{AB} = \frac{(n_A \bar{y}_A + n_B \bar{y}_B)}{(n_A + n_B)} \]  
(Eq. 4.18)

and \( n_A, n_B, n_{AB} = n_A + n_B \) are the numbers of points in cluster \( A \), cluster \( B \) and cluster \( AB \). The terms \( SSE_A, SSE_B, SSE_{AB} \) are the within-cluster sum of squares.

A generalization formula describing the different hierarchical clustering methods is proposed by Lance and Williams, (1967) and is having the following form (Equation 4.19):

\[ D(C_i \cup C_k, C_j) = \alpha_i D(C_i, C_j) + \alpha_j D(C_k, C_j) + \beta D(C_k, C_i) + \gamma \left| D(C_i, C_j) - D(C_k, C_j) \right| \]  
(Eq. 4.19)

where \( \alpha, \beta, \gamma \) are coefficients; \( D \) is distance function; \( C_j \) is the new cluster obtained from the joined of the clusters \( C_i \) and \( C_k \). The above formula (Equation 4.19) for different values of the
coefficients produces different hierarchical clustering methods as in Table 4.3 (Hardle and Simar, 2003; Goncalves et al., 2008), where in the below Table the $n_i$ stands for the number of objects that belong to cluster $C_i$.

**Table 4.3.** Computation of the different hierarchical clustering methods according to different coefficients values based on the Equation 4.19.

| Method          | $a_i$     | $a_j$     | $\beta$ | $\gamma$ |
|-----------------|-----------|-----------|----------|----------|
| Single linkage  | $1/2$     | $1/2$     | $0$      | $-1/2$   |
| Complete linkage| $1/2$     | $1/2$     | $0$      | $1/2$    |
| Average linkage | $n_i/(n_i+n_k)$ | $n_k/(n_i+n_k)$ | $0$ | $0$ |
| Centroid        | $n_i/(n_i+n_k)$ | $n_k/(n_i+n_k)$ | $-n_i n_k/(n_i+n_k)^2$ | $0$ |
| Ward            | $(n_i+n_j)/(n_i+n_k+n_j)$ | $(n_k+n_j)/(n_i+n_k+n_j)$ | $-n_j/(n_i+n_k+n_j)$ | $0$ |

The results of the hierarchical clustering methods can be visualized with the use of a graphical representation called dendrogram. In the case of *hierarchical* method, the groups’ number is unknown and the final structure of groups and subgroups is presented with the use of the dendrogram (Sivri et al., 2017). A dendrogram can display the observations, the sequence of clusters and the distances between the clusters and large distances indicate the clustering of heterogeneous groups (Hardle and Simar, 2003). Because the hierarchical clustering methods proceed until only one cluster to remain, the number of how many clusters to retain is not restrictive (Gore, 2000). As stated by Moskalik et al., (2014), the determination of the number of the clusters is a tricky problem and for that purpose several dozens of automated methods exist, however in the case that there are environmental pointers to the numbers of features investigated then they will be equal to the number of the clusters.
A widely used methodology for determining the number of clusters for the hierarchical clustering methods is calculating the Sneath's index of cluster significance. The Sneath’s index of cluster significance gives the clusters number with the use of a cutoff line for the dendrogram that is created after using a hierarchical clustering method. The stringent criterion of Sneath's index of cluster significance equals to \( \frac{2}{3} \) of \( D_{\text{max}} \) and the less restrictive criterion of Sneath's index of cluster significance equals to \( \frac{1}{3} \) of \( D_{\text{max}} \) (Astel et al., 2007; Kosiba et al., 2011; Oszust et al., 2014), where \( D_{\text{max}} \) is the maximum value of the bond distance \( D \) (usually the Euclidean).

For example, Kosiba et al. (2011) used hierarchical cluster analysis and a dendrogram in order to examine water properties and trophicity of five sampling sites in Poland aiming the restoration of the rare plant Aldrovanda Vesiculos. As it can be observed (based on Figure 4.2) the stringent criterion of Sneath's index of cluster significance separated the sampling site

![Figure 4.2](image_url)

**Figure 4.2.** Dendrogram of similarity and the cutoff lines that create groups, based on the stringent criterion and the less restrictive criterion of Sneath's index of cluster significance. The symbols are explained as: AO\(_i\); PR\(_i\); CH\(_i\); NO\(_i\); Nr\(_i\) the sampling sites and i the sampling number (taken by Kosiba et al., 2011)
into two groups, which differ in trophicity. The less restrictive criterion of Sneath's index of cluster significance separated the second group into three subgroups, based on their water parameters properties.

### 4.6. Multiple Linear Regression

Multiple Linear Regression (MLR) models are predicting a dependent variable from a set of independent variables. The mathematical equation (Equation 4.20) that describes the general MLR model with $Y$ the dependent variable and $X_1, X_2, \ldots, X_p$ the independent variables is given as below (Weiseberg, 2005):

$$ E(Y/X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \text{ (Eq. 4.20) } $$

where the symbol $X$ in $E(Y/X)$ means that we are conditioning on all the terms on the right site of the equation and the terms $\beta_i$ are constant parameters.

Despite the simple form of the MLR model, the multipredictor linear regression model is flexible and represents realistic how the average value of the outcome varies systematically in accordance with the predictors (Vittinghoff et al., 2005). An important drawback of the MLR models is that they can only simulate linear relationships, but they are widely used because of their simple structure with relatively few parameters (Gan et al., 2004).

The MLR models are widely used in water quality modelling and for the prediction of Chl-a levels. For the Chl-a levels prediction several modelling studies exist with good modelling results like the one of Cho et al., (2009). Some researchers applied more sophisticated modelling techniques for the application of the MLR model, like the combination of PCA and the MLR model method. The purpose of that technique is to simplify
the complexity of the relations between the independent variables. In that case before applying
the MLR method PCA was performed for the independent variables. In such an example
Camdevyren et al. (2005) predicted the Chl-a using 76,1% of the variation of biotic and
abiotic water quality parameters and with predictive success to be $R^2 = 90.8\%$, reducing by
that way the MLR model dimensions.

4.7. Studies combining ANN models with other modelling methodologies

ANNs are a powerful modelling tool, however sometimes the combination of ANNs with
some other modelling methodologies, e.g. PCA, can provide a solution for some modelling
issues that might appear. For example, the problem of computational complexity might result
into the ANN to not perform well enough, therefore the ANN model's dimensionality must be
reduced by removing any possible variables that increase the computational error of the ANN.
The phenomenon of spatial heterogeneity appears very often when the data set is taken from a
large geographical area. In that case the ANN model might not be able to simulate well
enough the whole data set, resulting to big simulation error for some inputs.

Another issue that might appear in environmental modelling studies is the phenomenon of
collinearity between the variables. As it is stated by Mela and Kopalle (2002) collinearity
might reduce the parameter estimates and that positive and negative correlation structures
have an asymmetric effect on variable omission bias. One of the most serious modelling
considerations is the omitted variables problem and it occurs when omitted variables affect the
relationship between the dependent variable and included explanatory variables (Leightner
and Inoue, 2012). Some proposed solutions for the problems mentioned above are found in
the literature. The way how ANNs water quality modelling studies and the methodology that
was used in order to resolve these problems are discussed below.
A very common tactic applied in ANN water quality modelling is to use PCA on the environmental parameters and to remove the less contributing ones. By that way dimension reduction is archived and the ANN model complexity is reduced. For example, Hadjisalomou et al. (2017) in their study performed PCA on data set of 9 water quality parameters (SRP; DIN; CO₃; HCO₃; pH; EC; WT; DO; SD) that are associated with Chl-a production. The PCA indicated that the carbonates were parameters of no importance and they were not introduced to the constructed ANN modelling Chl-a levels. A second ANN model that included the CO₃; HCO₃ as input parameters verified that the carbonates parameters were increasing computational complexity and produced output values with less accuracy (see Table 4.4).

**Table 4.4.** Evaluation of ANN performance for the two parameter sets based on the RMSE, root mean square error; MAE, absolute mean error and R, correlation coefficient (taken by Hadjisalomou et al., 2017)

| Model Inputs                              | R   | MSE  | MAE  |
|-------------------------------------------|-----|------|------|
| **Set 1:** SRP(t-1), DIN(t-1), pH(t-1), EC(t-1), WT(t-1), DO(t-1), SD(t-1) and Chl-a(t-1) | 0.959 | 28.46 | 15.75 |
| **Set 2:** SRP(t-1), DIN(t-1), pH(t-1), EC(t-1), WT(t-1), DO(t-1), SD(t-1) and Chl-a(t-1), CO₃(t-1), HCO₃(t-1) | 0.923 | 32.62 | 19.35 |
Another ANN modelling technique that is using PCA method in order to reduce ANN computational complexity is by introducing as inputs to the ANN the PCs instead of the parameters. In the study of Zounemat-Kermani (2014) the Chl-a levels are estimated with the use of ANN model and PCA is applied on 14 potential environmental parameters, resulting in eight groups of PCs that are used as possible input parameters. The ANN model results showed that the Chl-a levels were estimated with better accuracy when two PCs (summarizing a total of six of the initial parameters) were used as inputs, comparing with the results of simulated Chl-a when all the environmental parameters are used as inputs.

The problem of spatial heterogeneity might appear when modelling data sets from large lakes. In that case the created ANN might be inadequate to produce reliable results. In their study Huang et al. (2015) address the problem of spatial heterogeneity when modelling the Ch-a in the large lake Poyang -China- by applying the k-means algorithm. The 17 sampling sites were clustered into two clusters. Afterwards three different ANN models were constructed, two had as input variables data from only one of the resulted clusters and the other one ANN had as inputs data for the whole data set, where this process is explained analytically as in the graph below (Figure 4.3). The ANN models that studied the data for each cluster separately had better results than the ANN model that simulated Ch-a for the whole data set.

The phenomenon of collinearity between variables is problem that might appear during environmental modelling with ANNs. ANNs hybrid models (Fig. 4.4) are solving the problem of collinearity, because they are able to establish the relationship between a dependent variable (y) and the independent variables (x_i) with the advantage of variable reduction and zero correlation between the score values, avoiding by that way the problem of collinearity (Ul-Saufie et al., 2013).
Figure 4.3. Modelling procedure for the three constructed ANNs, where ANN_Upstream and ANN_Downstream models simulate data belong to different clusters. The ANN_Poyang simulates simulates data of the whole data set (taken by Huang et al., 2015).

Figure 4.4. The process of a hybrid model using PCA on the variables and afterwards the resulted PCs are used as inputs for the rest modeling process based on an ANN and MLR (taken by Ul-Saufie et al., 2013).
In their study Zhang et al. (2010) used a hybrid method that is using an ANN and principal component regression (PCR) in order to predict Chl-a and the basics of the procedure that was followed in that study are described below. The PCR is the combination of PCA and MLR method. PCs were used as explanatory variables for MLR in order to remove collinearity and the linear component was estimated by PCR. Then the ANN was used for estimating the model residual that was a nonlinear component. This hybrid methodology is based on computing the following equations:

\[ Y = G + N \]  \hspace{1cm} \text{(Eq. 4.21)}

\[ e = Y - \hat{G} \]  \hspace{1cm} \text{(Eq. 4.22)}

where \( G \) is a linear component and \( N \) a nonlinear component. The \( \hat{G} \) is estimated by applying different numbers of PCs (stepwise regression option) using the PCR method and the estimated value is symbolized as \( \hat{G} \). \( Y \) is symbolized the observed value and with \( e \) is symbolized the residual. The ANN is used for computing the residuals and the combined result of this hybrid methodology symbolized as \( \hat{Y} \) is given as below (Equation 4.23):

\[ \hat{Y} = \hat{G} + \hat{J} \]  \hspace{1cm} \text{(Eq. 4.23)}

where the term \( \hat{J} \) is the forecasted residual by ANN.
4. 8 ANN models vs other modelling methodologies

The modelling techniques analyzed above in most cases are producing good modelling results and for that reason are widely used. However, they are not suitable for some data sets and for that reason they can not reproduce reliable results. In that case it is observed that the ANNs can be applied and to give well simulated outputs. This is related to the fact that the ANN is advanced modelling technique with superior modelling abilities. Many modelling studies have observed the superiority of ANN models against other classical modelling techniques. In contrast with MLR models, the ANN models are not providing simple equations, but they also can easily quantify the contribution of each variable (Lek et al., 1996). Some representative examples of water quality modelling studies that are comparing the results of ANNs against the classical modelling techniques are given below.

For example, in their water quality modelling study Ozel et al. (2017) simulated the biological oxygen demand (BOD) parameter in river Bardin (Turkey) with the use of chemical oxygen demand (COD); suspended solids (SS); water temperature (WT), pH; electric conductivity (EC) parameters. Three different models were examined: the MLR model; the Multilayer Perceptron Neural Network (MLP); Radial Basis Neural network (RBANN). The models were evaluated with the use of the MAE; RMSE and determination coefficient ($R^2$) and the ANN models had better performance comparing with the MLR model (see Figure 4.5).
Many other modelling studies were contacted simulating the Chl-a parameter with both ANN and MLR methodologies and it was found that the ANN models had better performance than the MLR models (e.g. Karul et al., 1999; Karul and Soyupak, 2003; Malek et al., 2011; Karakaya et al., 2013; Rajaee and Boroumand, 2015). Some exemptions are cited in the

**Figure 4.5.** Simulation of the BOD parameter and comparison of the modelling results between the MLR model; MLP model; RBANN model (taken by Ozel et al., 2017).
literature, where the MLR models have better performance than the ANN models when simulating the Chl-a parameter. For example, in their study Merdun and Cinar (2010) simulated the Chl-a parameter with ANN and MLR models and found that the MLR performed slightly better than the ANN model, however it was noticed that the ANOVA test results showed that the difference was not statistically significant ($p<0.05$).

The superiority of the ANN models against MLR models is based on their inherent non-linearity and inter-connectivity of ANN together with its ability to learn and generalize information from complex or poorly understood systems (Gan et al., 2004). Besides the fact that the ANN models are better predictors than the MLR models, they are having the advantage of normality for their residuals and their independence from the variable to be predicted and the ability to produce good results with non-transformed data, in contrast with MLR models (Lek et al., 1996).

The phenomenon of dendrogram overcrowding is observed when dendrograms may grow to be so overloaded that leafs become overcrowded, while this drawback of dendrograms is increasing with sample size and may cause a problem of dendrogram readability (Kontaxaki et al.; 2010). In their study Sanchez-Martos et al. (2002) are using SOM model and cluster analysis in order to model the groundwater quality based on data set of 168 observations and 9 variables. During the cluster analysis application on the raw (initial) data set the phenomenon of dendrogram overcrowding is observed (see Figure 4.6), where it is difficult to identify groups containing samples that correspond to similar water quality. This problem is attributed to the large number of sampling points in the data set.
In contrast the SOM modelling procedure managed to represent each sampling point with an activation map of 9x9 neurons and the resulted dendrogram (see Figure 4.7) to be easily implemented. Therefore, the SOM method is producing results that are more easily to be implemented with the use of a dendrogram, comparing with the hierarchical clustering methodologies. According to Astel et al. (2007) both the cluster analysis algorithms and the SOM methodologies have the advantage of visualization ability, but the SOM has three more additional advantages:

I. SOM visualization is having the ability to present similarity between positive correlated variables but is also able to present similarity between negative correlated variables.

II. SOM visualization is able to detect outliers, data samples that don belong to a well-organized group.

III. The projection of variables similarity contains semi-quantitative information about the
distribution of a given parameter in the sampling locations space.

Figure 4.7. Dendrogram obtained obtained from the activation maps (taken by Sanchez-Martos et al., 2002)

Modelling studies that are using both PCA and the SOM methods are usually producing similar results (e.g. Samecka-Cymerman et al., 2007; An et al., 2016). However, it is observed that the SOM ability to classify environmental samples is more advanced than the PCA ability. In the study of Choi et al. (2014) the SOM and PCA were used in order to classify groundwater monitoring data in South Korea. Because the groundwater samples were widely dispersed, it was noticed that the PCA results were not satisfactory to classify groundwater and the loading values of the variables did not provide any meaningful information. However, the SOM managed to relate the groundwater samples with the variables and also to classify the groundwater samples into different groups.

In contrast with the traditional statistical methodologies that are not able to present well the information of large data sets, the SOM is suitable for visualization of high-dimensional data, because SOM is converting complex statistical relationships between data sets into simple geometric relationships on a low-dimensional display (Varbio et al., 2007). According to Lee and Scholz, (2006) the outstanding performance of the SOM model comparing with
other modelling techniques is related with the SOM's potential of clustering and classifying data, but also the SOM is ideal for detecting outliers because they are displaying in particular parts of the map without affecting the remaining parts and each outlier takes its place in one unit of the map.

Some other advantages of the SOM that are making it a superior modelling method comparing with the traditional statistical methodologies are: the ability to handle missing data; dealing with nonlinearities of the system; the SOM can be developed from data without requiring a mechanistic knowledge of the system; easily and quickly updated; noise reduction or handling irregular data; SOM provides an advanced visualization method (Cinar and Merdun, 2009; Vesanto and Alhoniemi, 2000). The SOM allows the visualization of association, while the sample units and the variables concentration can be displayed in the same figure making data interpretation easier (Giraudel and Lek, 2001). Based on these mentioned above the superiority of SOM is obvious, when comparing with other statistical methods.

4. 9. References

An, Y.; Zou, Z.; Li, R. Descriptive Characteristics of Surface Water Quality in Hong Kong by a Self-Organising Map. *Int J Environ Res Public Health*. 2016, 13(1). doi: 10.3390/ijerph13010115.

Astel, A.; Mazerski, J.; Polkowska, Z.; Namiesnik, J. Application of PCA and time series analysis in studies of precipitation in Tricity (Poland). *Advances in Environmental Research* 2004, 8, 337-349.
Astel, A.; Tsakovski, S.; Barbieri, P.; Simeonov, V. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research* **2007**, *41*, 4566-4578.

Basilevsky, A. *Statistical Factor Analysis and Related Methods*. Wiley: Canada, 1994.

Camdevyren, H.; Demyr, N.; Kanik, A.; Keskyn, S. Use of Principal Component scores in multiple linear regression models for prediction of *Chlorophyll-a* in reservoirs. *Ecological Modelling* **2005**, *181*, 581-589.

Chen, J.; Li, Fayun, L.; Fan, Z.; Wang, Y. Integrated Application of Multivariate Statistical Methods to Source Apportionment of Watercourses in the Liao River Basin, Northeast China. *Int. J. Environ. Res. Public Health* **2016**, *13*, 1035; doi:10.3390/ijerph13101035

Cho, K.H.; Kang, J.H.; Ki, S.J.; Park, Y.; Cha, M.S.; Kim, J.H. Determination of the optimal parameters in regression models for the prediction of *chlorophyll-a*: A case study of the Yeongsan Reservoir, Korea. *Science of the Total Environment* **2009**, *407*, 2536-2545.

Choi, B.-Y.; Yun, S.-T.; Kim, K.-H.; Kim, J.-W.; Kim, H. M.; Koh, Y.-K. Hydrochemical interpretation of South Korean groundwater monitoring data using Self-organizing Maps. *Journal of Geochemical Exploration* **2014**, *137*, 73-84.

Chu, H.J.; Liu, C.Y.; Wang, C.K. Identifying the Relationships between Water Quality and Land Cover Changes in the Tseng-Wen Reservoir Watershed of Taiwan. *Int. J. Environ. Res. Public Health* **2013**, *10*(2), 478-489; doi:10.3390/ijerph10020478.

Cinar, O.; Merdun, H. Application of an unsupervised artificial neural network technique to multivariant surface water quality data. *Ecol Res* **2009**, *24*, 163-173.

Duan, B.; Liu, F.; Zhang, W.; Zheng, H.; Zhang, Q. et al. Evaluation and Source Apportionment of Heavy Metals (HMs) in Sewage Sludge of Municipal Wastewater
Treatment Plants (WWTPs) in Shanxi, China. *Int. J. Environ. Res. Public Health* **2015**, 12(12), 15807-15818; doi:10.3390/ijerph121215022

Foster, J.; Barkus, E.; Yavorsky, C. *Understanding and using Advanced Statistics*. Sage publications: London, UK, 2006.

Gan, T. Y.; Kalinga, O.A.; Ohgushi, K.; Araki, H. Retrieving seawater turbidity from Landsat TM data by regressions and an artificial neural network. *International Journal of Remote Sensing* **2004**, 25, 4593-4615.

Giraudel, J. L.; Lek, S. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination. *Ecological Modelling* **2001**, 146, 329-339.

Goncalves, M.L.; Netto, M.L.A.; Costa, J.A.F.; Zullo Junior, J. An unsupervised method of classifying remotely sensed images using Kohonen self-organizing maps and agglomerative hierarchical clustering methods. *International Journal of Remote Sensing* **2008**, 29, 3171-3207.

Gore, P.A. Cluster Analysis. In *Handbook of Multivariate Statistics and Mathematical Modeling*; Tinsley, H.E.A.; Brown, S.D., Eds.; Academic Press: California, USA, 2000; pp. 297-322.

Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G.; Papastergiadou, E. Evaluating the contributing environmental parameters associated with eutrophication in a shallow lake by applying artificial neural networks techniques. *Fresenius Environmental Bulletin* **2017**, 26 (5), 3200-3208.

Hardle, W.; Simar, L. *Applied Multivariate Statistical Analysis*; Springer-Verlag: Berlin Heidelberg, Germany, 2003.
Hou, K.; Li, X.; Zhang, J. GIS Analysis of Changes in Ecological Vulnerability Using a SPCA Model in the Loess Plateau of Northern Shaanxi, China. *Int. J. Environ. Res. Public Health* **2015**, *12*, 4292-4305; doi:10.3390/ijerph120404292

Huang, J.; Gao, J.; Zhang, Y. Combination of artificial neural network and clustering techniques for predicting phytoplankton biomass of Lake Poyang, China. *Limnology* **2015**, *16*, 179-191.

Jakson, D.A. Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology* **1993**, *74*(8), 2204-2214.

Kagalou, I and Leonardos, I. Typology, classification and management issues of Greek lakes: implication of the Water Framework Directive (2000/60/EC). *Environ Monit Assess* **2009**, *150*, 469–484.

Karakaya, N.; Evrendilek, F.; Gungor, K.; Onal, D. Predicting Diel, Diurnal and Nocturnal Dynamics of Dissolved Oxygen and Chlorophyll-a Using Regression Models and Neural Networks. *Clean Soil Air Water* **2013**, *41*, 872–877. doi:10.1002/clen.201200683

Karul, C.; Soyupak, S.; Yurteri, C. Neural Networks as management tools in lakes. *Hydrobiologia* **1999**, *408/409*, 139-144.

Karul, C.; Soyupak, S. A Comparison between Neural Network Based and Multiple Regression Models for Chlorophyll-a Estimation. In: *Ecological Informatics*; Recknagel F., eds, Springer: Berlin, Heidelberg, Germany, 2003.

Kontaxaki, S.; Kokla, M.; Kavouras, M. Geo-Scape, a Granularity Depended Spatialization Tool for Visualizing Multidimensional Data Sets. *Geo-Spatial Information Science* **2010**, *13*, 275-284.
Kosiba, P.; Mroz, L.; Kaminski, R. Assessment of habitat conditions using self-organizing feature maps for reintroduction/introduction of *Aldrovanda Vesiculosa* L. in Poland. *Acta Societatis Botanicorum Poloniae* **2011, 80**, 139-148.

Kura, N.U.; Ramli, M.F.; Sulaiman, W.N.; Ibrahim, S.; Aris, A.Z.; Mustapha, A. Evaluation of Factors Influencing the Groundwater Chemistry in a Small Tropical Island of Malaysia. *Int. J. Environ. Res. Public Health* **2013, 10**(5), 1861-1881; doi:10.3390/ijerph10051861.

Lance, G.N. and Williams, W.T.A., General theory of classificatory sorting strategies: 1. Hierarchical Systems, *Computer Journal* **1967**, 373-380.

Lee, B.-H., Scholz, M. A comparative study: prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks. *Water, Air, and Soil Pollution* **2006**, **174**, 279-301.

Leightner, J.; Inoue, T. Solving the Omitted Variables Problem of Regression Analysis Using the Relative Vertical Position of Observations. *Advances in Decision Sciences* **2012**, doi: 10.1155/2012/728980.

Lek, S.; Delacoste, M.; Baran, P.; Dimopoulos, I.; Lauga, J.; Aulagnier, S. Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling* **1996, 90**, 39-52.

Malek, S.; Ahmad, S. M. S.; Singh, S. K. K.; Milow, P.; Salleh, A. Assessment of predictive models for chlorophyll-a concentration of a tropical lake. *BMC Bioinf.* **2011, 12**(Suppl. 13), S12.10.1186/1471-2105-12-S13-S12

Marques de Sa, J.P. *Applied Statistics Using SPSS, STATISTICA, MATLAB and R*, 2nd ed.; Springer-Verlag: Berlin Heidelberg, Germany, 2007.

Mela, C.; Kopalle, P. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Applied Economics* **2002, 34**, 667-677.
Merdun, H.; Cinar, O. Artificial Neural Network and Regression Techniques in modelling surface water quality. *Environment Protection Engineering* **2010**, *36*, 95-109.

Millard, P.S. and Neerchal, K.N. *Environmental Statistics with S-PLUS*. CRC Press: Boca Raton, FL, USA, 2000.

Moskalik, M.; Tegowski, J.; Grabowiecki, P.; Zulichowska, M. Principal Component and Cluster Analysis for determining diversification of bottom morphology based on bathymetric profiles from Brepollen (Hornsund, Spitsbergen). *Oceanologia* **2014**, *56*, 59-84.

Mukaka, M.M. A guide to appropriate use of Correlation coefficient in medical research. *Malawi Med J.* **2012**, *24*(3): 69–71.

Oszust, K.; Frac, M.; Gryta, A.; Bilinska, N. The Influence of Ecological and Conventional Plant Production Systems on Soil Microbial Quality under Hops (*Humulus lupulus*). *Int J Mol Sci.* **2014** Jun; 15, 9907–9923. doi:10.3390/ijms15069907

Ozel, H. U.; Gemici, B. T.; Ozel, H. B.; Gemici, E. Determination of water quality and estimation of monthly biological oxygen demand (BOD) using by different artificial neural networks models in the Bartin River. *Fresenius Environmental Bulletin* **2017**, *26*, 5465-5476.

Papatheodorou, G.; Hotos, G.; Geraga, M.; Avramidou, D.; Vorinakis, T. Heavy metals concentrations in sediments of Klisova lagoon (Southeast Mesolonghi-Aetilikon Lagoon complex), W. Greece. *Fresenius Environmental Bulletin* **2002**, *11*, 951-956.

Rajaee, T.; Boroumand, A. Forecasting of chlorophyll-a concentrations in South San Francisco Bay using five different models. *Applied Ocean Research* **2015**, *53*, 208-217.

Reimann, C.; Filzmoser, P.; Garrett, R.; Dutter, R. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*; Wiley: West Sussex, UK, 2008.
Rencher, A.C. *Methods of Multivariate Analysis*. 2nd ed.; Wiley Interscience: New York, USA, 2002.

Samecka-Cymerman, A.; Stankiewicz, A.; Kolon, K.; Kempers, A. J. Self-organizing feature map (neural networks) as a tool in classification of the relations between chemical composition of aquatic bryophytes and types of streambeds in the Tatra national park in Poland. *Chemosphere* **2007**, *67*, 954-960. Doi: 10.1016/j.chemosphere.2006.11.001

Sanchez-Martos, F.; Aguilera, P.; Garrido-Frenich, A.; Torres, J.; Pulido-Bosch, A. Assessment of Groundwater Quality by Means of Self-organizing Maps: Application in a Semiarid Area. *Environmental Management* **2002**, *30*, 716-726.

Sarstedt, M and Mooi, E. A Concise Guide to Market Research. 2nd ed.; Springer-Verlag: Berlin Heidelberg, Germany, 2014.

Sivri, N.; Cilingirturk, A.M.; Seker, D.Z.; Imamoglu, Z.; Ucan, O.N. Prediction of Water Consumption in Istanbul by means of statistical forecasting models & Geographig Information Systems (GIS). *Fresenius Environmental Bulletin* **2017**, *26*, 375-382.

Stefanidis, K. and Papastergiadou, E. Relationships between lake morphometry, water quality, and aquatic macrophytes, in Greek Lakes. *Fresenius Environmental Bulletin* **2012**, *21*, 3018-3026.

Tan, P.-N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*. Addison-Wesley: Boston, MA, USA, 2006.

Ul-Saufie, A. Z.; Yahaya, A. S.; Ramli, N. A.; Rosaid, N.; Hamid, H. A. Future daily PM$_{10}$ concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). *Atmospheric Environment* **2013**, *77*, 621-630.
Varbiro, G.; Acs, E.; Borics, G.; Erces, K.; Feher, G. et al. Use of Self-organizing Maps (SOM) for characterization of riverine phytoplankton associations in Hungary. *Arch. Hydrobiol. Suppl.* **2007**, *161*, 383-394.

Vesanto, J.; Alhoniemi, E. Clustering of the Self-organizing Map. *IEEE Transactions on Neural Networks* **2000**, *11*, 586-600.

Vittinghoff, E.; Shiboski, S.; Glidden, D.; McCulloch, C. *Regression Methods in Biostatistics: linear, logistic, survival, and repeated measures models*. Springer: New York, USA, 2005.

Weisberg, S. *Applied Linear Regression*. Wiley-: New Jersey, USA, 2005.

Zhang, W.; Li, H.; Sun, D.; Zhou, L. A Statistical Assessment of the Impact of Agricultural Land Use Intensity on Regional Surface Water Quality at Multiple Scales. *Int. J. Environ. Res. Public Health* **2012**, *9*(11), 4170-4186; doi:10.3390/ijerph9114170.

Zhang, X.; Wang, Q.; Yu, M.; Wu, J. Combining Principals Component Regression and Artificial Neural Network to Predict Chlorophyll-a Concentration of Yuqiao Reservoir's Outflow. *Trans. Tianjin Univ.* **2010**, *16*, 467-472.

Zounemat-Kermani, M. Principal Component Analysis (PCA) for estimating chlorophyll concentration using forward and generalized regression neural networks. *Applied Artificial Intelligence* **2014**, *28*, 16-29.
5. Application 1: Assessing the Contribution of the Environmental Parameters to Eutrophication with the Use of the “PaD” and “PaD2” Methods in a Hypereutrophic Lake

5. 1. Introduction

Over the past few decades eutrophication has emerged as a serious problem affecting the water quality of many lakes worldwide, mainly as a result of increased nutrient loadings related to human activities (Ferreira et al., 2007). Eutrophication is related with phytoplankton overproduction and has as a common proxy: the chlorophyll-\( \alpha \) (Chl-\( \alpha \)) parameter. Eutrophic conditions in a water column may lead to a Harmful Algal Bloom (HAB) event (Dyhrman, 2008). Three different categories of HABs can be distinguished: toxic algae; potentially toxic algae; and high-biomass blooms, which are called “red tides” (Ferreira et al, 2011).

In the case of toxic HABs events, toxin-producing algal species are involved. In freshwater lakes most of these toxic species are cyanobacteria, formally called blue-green algae (Ha and Pflugmacher, 2013). Because of their toxin production, cyanobacteria have been characterized as potential key hazardous pollutants, by the European Water Framework Directive (2000) (2000/60/EC) (Hoeger et al., 2005). Aquatic organisms may suffer from poisoning because of cyanotoxins released from cyanobacteria when a HAB occurs (Liu et al., 2011). The role of cyanotoxins in recreational waters and their effect on human health have started to be under examination the recent years (Koreiviene et al., 2014). It is reported that
about 60,000 intoxication incidents, with an overall mortality rate 1.5%, take place per year globally because of algal toxins (Ferrante et al., 2013).

High biomass blooms are associated with hypoxia and anoxia, with partial or even total oxygen depletion in the water (Gubelit and Berezina, 2010). In hypoxic waters aquatic ecosystems and public health is under serious threat. Besides the negative impact on public health and aquatic organisms, high biomass blooms are responsible for a series of other problems affecting water quality, like bad odor and algal scum formation (Atoui et al., 2013).

The economic cost of these eutrophication related problems is huge. As it is stated by Dodds et al. (2009) the economic cost of eutrophication in U.S. freshwaters reached $2.2 billion annually, with the greatest economic loses related with lakefront property values and recreational use.

Understanding the links between Chl-a and the associated environmental parameters is the first critical step for eutrophication management. The linkage between eutrophication and environmental parameters is often non-linear and complex (Paerl, 2006), so the simple linear structure of a regression model cannot be applied and other mathematical methods are more suitable for lake modelling. Artificial Neural Networks (ANNs) have proven to be a powerful tool in lake eutrophication assessment (Xu et al., 2001), because of their ability to model phenomena involving non-linear and complex data (Lek and Guegan, 1999).

The aim of this study was the creation of an ANN model that would simulate the Chl-a and the examination of the impact of each environmental parameter associated with eutrophication. The necessity for evaluating in detail the role of each environmental parameter arose because despite several restoration attempts, Lake Pamvotis remains a hypereutrophic lake. The contribution of each parameter was calculated with the use of the Partial Derivatives ("PaD") method. The coupled effect of the parameters were calculated with the use of the “PaD2” method, in an attempt to have a more thorough look into eutrophication process and to
examine the synergistic effects between the environmental parameters. To our knowledge this is the first modelling study where the “PaD2” algorithm is applied in order to examine the synergistic effect of environmental parameters on the simulated Chl-α.

The “Pad2” algorithm is a two-way interaction sensitivity method designed for ANNs that provides information regarding the impact of the several combinations of input parameters to the modeled output. In that way the synergistic effects of the environmental parameters is examined and useful conclusions regarding the trophic production can be extracted. By applying the “Pad2” algorithm the input environmental parameters synergy mechanism is revealed and the environmental parameters’ interactions are evaluated by the ANN model. Therefore this ANN modelling study can be considered an advanced management tool compared with other ANN modelling studies that examine only the one-way sensitivity of environmental parameters. The application of these two ANN model sensitivity methods produced interesting results that can act as an advisory tool for any future restoration attempts.

5.2. Materials and Methods

5.2.1. Study Area and Data Collection

Lake Pamvotis is a shallow Mediterranean lake located in northwestern Greece (Figure 5.1) with mean depth 4.3 m and maximum depth 7.5 m, raising about 470 m above sea level (Papatheodorou et al., 2006). It occupies an area of about 22.8 km² and is located next to the city of Ioannina. Pamvotis Lake has been listed among the Natural Special Conservation Areas under the European Community Council Directive on the conservation of natural habitats and of wild fauna and flora (Habitats Directive, EC, 92/43).

Lake Pamvotis can be consider as a “closed” hydrological system, as it has no natural surface outflows and is recharged by karstic springs (Papatheodorou et al., 2006; Kagalou et al., 2008). During the last several decades the lake has become more eutrophic due to impacts
from agricultural activities, livestock, nearby industrial units, discharges of domestic sewage, irrigation, sediment deposit and introduction of alien fishes (Gkelis et al., 2014).

![Satellite map of the study area](image)

**Figure 5.1.** Satellite map of the study area, where Ioannina city is located northwestern. The urban area of Ioannina city is observed in the west side of the lake.

A detailed overview of the key changes in the trophic state of the Lake Pamvotis can be found in the articles of Papatheodorou et al. (2006); Kagalou et al. (2008); Papastergiadou et al. (2010) and a summary of their recorded data is given below. The nutrient and Chl-\(a\) concentrations, presented for three different periods from 1985 to 2005, indicate an eutrophic-hypertrophic profile of the trophic state of the lake. Specifically, during the 1985–1988 period the mean annual Chl-\(a\) concentration was 76.91 mg/m\(^3\), demonstrating a highly eutrophic to hypertrophic classification of the trophic state. After the sewage diversion that occurred in early 1990s the annual mean of Chl-\(a\) decreased to 21.21 mg/m\(^3\), still showing strong signs of eutrophication. However, from samplings carried out in 2004–2005 it was demonstrated that Chl-\(a\) levels increased significantly, with an annual mean of 79.23 mg/m\(^3\) suggesting a shift of the trophic state towards hypertrophic conditions.

In previous years and particularly in 2005, massive cyanobacterial surface blooms were recorded from early summer to late autumn (Kagalou et al., 2008). More recently, Gkelis et al.
(2014) assessed the occurrence of cyanobacteria blooms in Lake Pamvotis and highlighted the co-occurrence of more than one cyanotoxins during the warm period and the potential risk for human health. These findings clearly suggest that the trophic status of the lake remains highly eutrophic despite the management practices that have been applied the last decades.

In an attempt to improve the lake trophic status, in 1986 Lake Pamvotis was stocked with several native fish species and the exotic planktivorus *Ctenopharyngodon idella*. The result of this intervention is debatable, since it led to great decline of submerged vegetation (Kagalou et al., 2003; Stafanidis and Papastergiadou, 2007). In 1992 the operation of the sewage treatment unit and the consequent discharge of the effluents into Lapsista channel began (Papastergiaodou et al., 2010).

The water quality parameters were collected during the monitoring period from June 2004–August 2005 on a monthly basis. Eleven monitoring stations inside Lake Pamvotis were used for the data collection. The measured parameters were SRP, DIN species (nitrite, nitrate, ammonium), carbonate (CO$_3$), bicarbonate (HCO$_3$), pH, electric conductivity (EC), water temperature (WT), dissolved oxygen (DO), Secchi disk (SD) and chlorophyll a (Chl-a).

### 5.2.2. ANNs Methodologies

Multi-layer feedforward ANNs are a very popular category of ANNs, mainly because they are suitable for function approximation. Feedforward ANNs have acyclic topology with no feedback loops among their layers and are used for approximating non-linear mappings among inputs and outputs (Hu and Hwang, 2002). A feedforward ANN has at least three layers and each layer consists of neurons. The first layer is the input layer that imports the input parameters to the network, one or more hidden layers and the output layer that gives the final result. For every neuron there is synaptic weight that connects it with every neuron of the next layer. The weighted sum of a neuron’s inputs produces an output with the use of a
transfer function (Kuo et al., 2007). The relationship that gives the ANN output from the $j$th neuron ($o_j$) as given by Kazemi Yazdi and Scholz (2010) has the form:

$$o_j = f(u_j) \quad \text{(Eq. 5.1)}$$

and:

$$u_j = \sum w_{ij}x_i + z_j \quad \text{(Eq. 5.2)}$$

where $f$ is a transfer function, $x_i$ is the input from it neuron belonging to the immediate previous layer, $w_{ij}$ is the synaptic weight that connects $x_i$ with the $j$th neuron and $z_j$ is a bias term.

The most widely used transfer functions are the linear transfer function, the log-sigmoid transfer function and the tangential sigmoid transfer function. ANNs are trained with the use of a learning algorithm. Feedforward ANNs are usually trained with the use of a back-propagation algorithm. During the learning procedure the synaptic weights are adjusted so they minimize an error function, usually the mean square difference between the predicted and the given output (Salami Shahid, 2016; Wang et al., 2016).

The number of hidden neurons has a major role in the good performance of an ANN. With a small number of hidden neurons the ANN will not be able to produce good results. Meanwhile for a big number of hidden neurons the phenomenon of overfitting occurs (Karul et al., 1999). Overfitting exists when the error for the training set is small, but the ANN has large errors for new data. In that case the ANN hasn’t learned to generalize to new situations (Demuth et al., 2007). To avoid overfitting many empirical rules that suggest the maximum number of hidden neurons have been proposed. One such practical rule is the one proposed by Maier et al. (1998):

$$N'' = 2N' + 1 \quad \text{(Eq. 5.3)}$$
and:

\[ N^H \leq 2N^{TR}/(N^I + 1) \]  \hspace{1cm} (Eq. 5.4)

where \( N^H \) is the number of hidden layer neurons, \( N^I \) the number of inputs and \( N^{TR} \) the number of training samples. The maximum \( N^H \) must be the smallest number calculated by those two equations. The determination of the optimal number of hidden neurons and layers is decided after a trial and error procedure (Ozesmi et al, 2006). In order to prevent an ANN’s overtraining early stopping and regularization methods are used. In early stopping the data set is divided into three subsets: training set, validation set and test set. The network training stops when the errors for the validation set begin to rise, indicating that the network had begun to overfit the data (Karul et al, 2000). The regularization method involves modification of the network performance function that is the sum of squares of the network errors in the training set. By using the modified performance function will cause the network to have smaller weights and biases, and this will force the network response to be smoother and less likely to overfit (Demuth et al., 2007).

The Partial Derivatives (“PaD”) method, presented by Dimopoulos et al. (1995), is a technique for measuring the ANN’s sensitivity regarding its input parameters and is a single parameter interaction method. This method has two parts. As analyzed by Gevrey et al. (2003) the partial derivatives of the output for small changes of each input are computed and then the relative contribution of each input variable is classified. The sensitivity of the ANN’s output for the input variable \( x_i \) is symbolized as \( SSDi \) and according to Dimopoulos et al. (1995) is the sum of the squared partial derivatives obtained per input variable and has the following form:

\[ d_{ij} = S \sum_{h=1}^{nh} w_{h\mu} I_{y_h}(1 - I_{y_h}) w_{ih} \]  \hspace{1cm} (Eq. 5.5)
where $N$ is the observations number of the data set; $d_{ji}$ the partial derivatives of the output $y_j$ with respect to input $x_j$. The equation for $SSD_i$ is given below:

$$SSD_i = \sum_{j=1}^{N} (d_{ji})^2$$

(Eq. 5.6)

where $nh$ the neurons number of the hidden layer; $I_{hj}$ the output of the $h$-th hidden neuron for the $j$-th input; $w_{ih}$ the weight connecting the $i$-th input neuron and the $h$-th hidden neuron; $w_{ho}$ the weight connecting the output and the $h$-th hidden neuron; $S_j$ is the derivative of the output $y_j$ with respect to input $x_j$.

The “PaD2” method examines the synergistic effects between the input parameters and is considered a two-way interaction sensitivity analysis method. The application of the “PaD2” method is very similar with the application of the “PaD” method. According to Gevrey et al. (2006) the first part of the “PaD2” method that gives the partial derivatives of the output $y_j$ with respect to inputs $x_j$ and $x_j$ is given by the following equation:

$$d_{ji} = \left[ s_j \sum_{h=1}^{nh} w_{1h} w_{ho} I_{hj} (1 - I_{hj}) \right] \left[ \sum_{h=1}^{nh} w_{2h} w_{ho} I_{hj} (1 - I_{hj}) \right]$$

(Eq. 5.7)

where the symbols were explained above, except $s_j$ that is the second derivative of the output neuron with respect to its input (Dimopoulos et al., 1995; van Maanen et al., 2010). The relative contribution of the coupled variables to the ANN is the sum of squared partial derivatives of the coupled variables and is calculated as below:

$$SSD_{12} = \sum_{j=1}^{N} (d_{j12})^2$$

(Eq. 5.8)
5.2.3. ANN Model Development

Using Principal Components Analysis (PCA) the selection of ANN’s input parameters is done (Zounemat-Kermani, 2014; Zhang et al., 2015). PCA is often combined with ANN modelling, because by that way dimension reduction is enabled and the model’s computational complexity is reduced, so the possibility of a model’s misconvergence and poor accuracy is eliminated (Muttil and Chau, 2007). Based on PCA results the carbonate and bicarbonate parameters were excluded, since they contribute less than 2% to the total variation of the data set. The resulting input parameters are SRP, DIN, pH, EC, WT, DO and SD and the model output is Chl-\(a\). In eutrophication modelling with ANNs it is a common practice the Chl-\(a\) to be log transformed and then the data set to be normalized (Scadi and Harding, 1999), as in that way the network performance is improved. After network training the data is set back to the initial form.

The method of regularization was used for avoiding model’s overfitting, for the regularization method the data set is divided into training set and test set. Therefore the data set consisting of 161 samples was divided into two subsets. The training set containing 80% of the data and the test set 20%. The ANN’s training algorithm was the Levenberg-Marquardt (LM) algorithm, because the LM algorithm has fast learning speed and is the best training algorithm among the rest variations of back-propagation algorithms for medium-sized networks (Zhang et al., 2015; Tota-Maharaj and Scholz, 2012). The sigmoid transfer function having the form:

\[
f(x) = \frac{1}{1 + e^{-x}} \quad \text{(Eq. 5.9)}
\]

was used as the activation function between the layers and by that way the non-linearity was imported into the model.
With the use of MatLab software, the ANN was simulated several times with different numbers of neurons in the hidden layer to find the optimal number of neurons as proposed by Sener et al. (2012), always keeping in mind the restrictions of the maximum number of neurons as given by the Equations (5.3) and (5.4). The optimal neurons number was 10 neurons for the hidden layer, seven neurons for the input layer and one neuron for the output layer. The resulting ANN optimal topology is thus 7-10-1, based in the form L1-H1-L2, where L1 is the number of neurons in the input layer, H1 the number of neurons in the hidden layer and L2 the number of neurons in output layer. The best ANN’s model performance was calculated for the test set using the coefficient of determination ($R^2$) and the Absolute Relative Error (RE) since they are commonly used statistical variables for evaluating model performance (Cui et al., 2016). The equations that give $R^2$and RE are given respectively as below:

$$R^2 = \left[ \frac{\sum (o_i - \bar{o}_i)(s_i - \bar{s}_i)}{\sqrt{\sum (o_i - \bar{o}_i)^2 \sum (s_i - \bar{s}_i)^2}} \right]^2 \quad \text{(Eq. 5.10)}$$

$$RE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{s_i - \bar{o}_i}{o_i} \right| \quad \text{(Eq. 5.11)}$$

where $o_i$ is the observed value; $\bar{o}_i$ is the average observed value; $s_i$ the simulated value; $\bar{s}_i$ the average simulated value and $N$ the observations number. The ANN model performance results for all the data subsets are presented in Table 5.1.

**Table 5.1.** ANN’s performance for the training, test and whole set based on the RE, Absolute Relative Error (RE) and $R^2$, coefficient of determination.
5.3. Results and Discussion

5.3.1. ANN’s Simulation Results

One-way sensitivity analysis was carried out in order to examine how strong the effect of each input parameter on the simulated output is. The “PaD” method was preferred among other ANN sensitivity techniques, since the “PaD” method is the most stable (Dimopoulos et al., 1995). The relative importance of input variables to the ANN was found to have the following order of contribution according the results of the “PaD” method: WT, SRP, SD, pH, DO, EC, DIN (Figure 2). The “PaD” method has the ability to discriminate the input variables into minor and major contributing environmental parameters (Dedecker et al., 2005). In our case the input parameters WT, SRP, SD, pH, DO have a major relative importance, while the input parameters EC, DIN have minor relative importance to the simulated Chl-α.

![Relative contribution of the input parameters to the ANN model with the use of Partial Derivatives (“PaD”) sensitivity method.](image)

**Figure 5.2.** Relative contribution of the input parameters to the ANN model with the use of Partial Derivatives (“PaD”) sensitivity method.
The “PaD2” algorithm was used to calculate the synergistic effect of paired input parameters on the modeled Chl-\(a\). For that purpose 21 combinations of paired input parameters were created. The evaluation of the impact for a small change of these paired parameters to the ANN was done with the application of the “PaD2” algorithm. The relative contribution of these interactions are given in Figure 5.3.

![Figure 5.3](image)

**Figure 5.3.** Relative importance of the paired input parameters to the ANN with the use of “PaD2” sensitivity method.

The sum of the five most influential paired parameters (WT-SRP, SD-WT, WT-pH, SD-pH, pH-SRP) described up to 70% of the ANN model sensitivity. The graphical representation of the most influential parameters enables us to produce useful results regarding the parameters’ interactions. The WT-SRP (Figure 5.4) is the most influential interaction and has a relative importance of almost 22%. As it is graphically observed the Chl-\(a\) gradients start to rise for high SRP concentrations combined with high WT. Meanwhile the same high WT values combined with lower SRP levels reveal that Chl-\(a\) gradients have the tendency to decrease. Therefore it is obvious that a decrease in SRP levels would lead to a decrease to Chl-\(a\) values, even during the hot months that are associated with high algal production (Paerl and Huisman, 2008).
Figure 5.4. Partial Derivative (D) of the ANN response to the paired interaction of the soluble reactive phosphorus (SRP) and water temperature (WT) variables.

Regarding the SD-WT interaction (Figure 5.5) it is noticed that the Chl-$a$ derivative starts to rise in an area of low WT values and medium-high SD values, but no clear conclusions can be reached. The ANN model managed to interpolate well the WT-pH interaction (Figure 5.6). For elevated pH and high WT there is a tendency for the Chl-$a$ gradient to rise. In that way the high pH levels observed related with increased algal activity and high WT that stimulates algal production, are associated. This is in agreement with a study of Trolle et al. (2012) which noted that high Chl-$a$ values are linked with high pH, especially during summer. The SD-pH combination (Figure 5.7) gives negative Chl-$a$ derivatives. This behavior can be explained by the fact that small increase of SD values due to application of the “PaD2” algorithm results in contrary behavior regarding Chl-$a$ production. The SD is negatively correlated with Chl-$a$ levels (LaBounty, 2008). The pH-SRP combination (Figure 5.8) although it follows no specific pattern, shows that for high SRP and pH levels the Chl-$a$ gradients take their highest values and for lower SRP levels the derivatives have smaller values, so the ANN managed to correctly connect the SRP effect on Chl-$a$ and to associate it with high pH because of increased algal production.
**Figure 5.5.** Partial Derivative (D) of the ANN response to the paired interaction of the secchi disk (SD) and water temperature (WT) variables.

**Figure 5.6.** Partial Derivative (D) of the ANN response to the paired interaction of the pH and water temperature (WT) variables.

**Figure 5.7.** Partial Derivative (D) of the ANN response to the paired interaction of the pH and secchi disk (SD) variables.
Figure 5.8. Partial Derivative (D) of the ANN response to the paired interaction of the pH and soluble reactive phosphorus (SRP) variables.

The SRP-DIN paired parameter was calculated to have no major relative importance on the simulated Chl-\(a\). Despite this finding the impact of DIN on Chl-\(a\) production should not be underestimated, because DIN high levels are associated with continued serious eutrophication problems caused by non-N\(_2\) fixing cyanobacterial blooms (Paerl et al., 2011). Additionally all the ANN simulations without the DIN as an input parameter had worse performance than the ANN simulations with the DIN as an input parameter, verifying the importance of DIN on the Chl-\(a\) in a mathematical way. For these reasons a graphical representation of the SRP-DIN (Figure 5.9) interaction is also presented, in order to exam the synergistic effect on Chl-\(a\) behavior.

Figure 5.9. Partial Derivative (D) of the ANN response to the paired interaction of the soluble reactive phosphorus (SRP) and dissolved inorganic nitrogen (DIN) variables.
The synergistic effect of the SRP-DIN follows a clear pattern regarding the Chl-\(a\) gradient. For high SRP and DIN values the Chl-\(a\) derivative has the bigger increase and as the SRP and DIN values decrease the Chl-\(a\) derivative tends to decrease accordingly. The reduction of DIN or SRP gives a reduction for Chl-\(a\) and the combined reduction of nutrients yields to even higher reduction (Sagrario et al.,2005), something that is verified by the “PaD2” algorithm findings.

5.3.2. Implications for Management and Restoration

This ANN modelling study was aimed at evaluating the influence of each environmental parameter on the simulated Chl-\(a\). A need for careful examination of the environmental parameters resulted, since it was observed that despite several restoration measures that were applied the last decades to Lake Pamvotis, no lasting water quality improvement was achieved. The impact magnitude of the parameters and the way that they interact with other parameters regarding algal production needed to be investigated. For that reason the “PaD” and the “PaD2” algorithms were used, producing interesting findings about the parameters.

An ANN model was chosen for this modelling study instead of any other modelling technique because of the ANNs’ ability to model complex systems, making them ideal for lake modelling. The trophic function of a lake can be modeled with the use of an ANN usually with a good correlation, as in this modelling study. Therefore the created ANN can be considered as a reliable predictor of Chl-\(a\) providing mathematical trustworthy relationships between the modeled Chl-\(a\) and the environmental parameters. Sensitivity analysis algorithms are the linkages between the Chl-\(a\) and the environmental parameters’ interactions. Furthermore the “PaD2” method enables us to assess the synergism of coupled parameters on the simulated Chl-\(a\), providing us an advanced management tool comparing with other ANN modelling
studies that deal only with one-way parameter sensitivity analysis. Interesting trends between variables synergistic effect and the simulated Chl-a were revealed, like the WT synergism with the SRP that makes restoration efforts more difficult.

Some of the input parameters were found to be more influential than others; however the importance of each input parameter must not be underestimated. All the ANN model’s input parameters were found to have an effect on the simulated Chl-a and an omission of any of the input parameters as an input would lead to poor ANN model performance. For that reason it was decided to pay attention to the DIN as well, even though it was found to have small relative importance for the Chl-a. The increased WT related with climatic change increases the risk for a potentially toxic cyanobacterial bloom in lakes with high nitrogen concentration (Jeppesen et al., 2011). The high relative importance of the SRP revealed that it is the key nutrient for Lake Pamvotis and has a significant role in the algal production. In a relevant study of Lake Pamvotis Papatheodorou et al. (2006) also found that the key nutrient is the SRP. The combined nutrients “synergistic” effect was pronounced in summer and fall, when algal production is highest (LaBounty, 2008). For that reason lake modelling studies that examine the nutrients reduction effect were carried out, like the one of Mateus et al. (2014). Our modelling study confirms these results, since the paired SRP-DIN parameter showed that a decrease of nutrients would lead to a Chl-a decrease. The “PaD2” algorithm findings regarding the nutrients behavior could act as a practical management guide tool for decision makers.

Another interesting finding was that the most important parameter was the WT, confirming that way Kagalou et al. (2008) theory that the effects of climate change for Lake Pamvotis may be greater than those caused by anthropogenic pressure. It is generally accepted that temperature increase leads to more intense eutrophication symptoms (Feuchtmayr et al., 2009; Moss et al., 2011). Keeping in mind the results of the SRP-DIN synergistic effect and
the major role of WT, then the necessity for extra nutrient reduction is obvious because the more eutrophic a lake is then the effect of temperature is greater on algal production (Ye et al., 2011). As it is stated by Papastergiadou et al. (2010) Lake Pamvotis is unlikely to switch to clear water conditions and has a tendency for eutrophic conditions, but improvement of water quality can be achieved through specific management practices like nutrient reduction from the catchment area and the sediment, increasing the flushing rate by increasing the karstic springs’ discharge by redverting the springs located on the northern shore into the lake, and by biomanipulation practices to minimize the population of the allochthonous benthivorous species. The Lake Pamvotis natural tendency for eutrophication can be partly explained by the paired SRP-WT parameter computational results. Climatic change favors the phosphate release from anoxic sediments and increases phosphate internal loading, especially during hot months (Mooij et al., 2009). Prolonged and intensified eutrophication might be related with an internal nutrient supply in the water body, even if external input no longer exists (Guneralp and Barlas, 2003). The climatic change and the associated WT increase must be taken into consideration for any restoration measures, since Lake Pamvotis is highly susceptible to increased WT effects.

For Lake Pamvotis other environmental factors that affect eutrophication should also be examined. For example modelling studies that examine the role of fish, macrophytes and zooplankton in lake restoration have been initiated. Models dealing with other restoration techniques like biomanipulation could be developed in order to provide alternative eutrophication management options. Sagehashi et al. (2001) created such a model for a shallow basin that combines as parameters planktivorous fish, three types of zooplankton, two types of algae and nutrients. Based on that model they simulated some restoration methods like biomanipulation, dredging and nutrient loading reduction and the analog effect on water quality. It is suggested as this ANN modelling study of Lake Pamvotis should be updated with
new extra environmental parameters that would include long-term monitored data like the ones mentioned above, e.g., planktivorous fish and zooplankton density.

5.4. Conclusions

Lake algal production is considered a very complex issue and the relationships with the associated environmental parameters are difficult to examine. However ANNs are ideal for lake modelling because of their ability to model non-linear situations. A small change of an input variable produces the sensitivity result of the ANN and the way the modeled output responds to that change. With the use of the “PaD” and “PaD2” methods the one-way sensitivity and the two-way sensitivity are calculated accordingly, revealing in that way how Chl-α responds to changes of the environmental variables. The results given by the “PaD” and “PaD2” methods are very important for understanding a lake function and how each environmental variable affects the algal production. Especially for eutrophic and hypertrophic lakes, these findings have the role of advisory restoration tools. For example by using the “PaD” method the key nutrient is found. Also with the use of “PaD2” method it is examined if the restoration measures must follow a one nutrient control policy or combined nutrient reduction. Meanwhile the synergistic effect between the environmental parameters can act as a warning tool for hypertrophic lakes, since the “PaD2” method can reveal if for the hot months and for certain nutrient levels a HAB event might occur.

The ANN managed successfully to simulate the Chl-α levels with a good correlation so the ANN model can be considered as a reliable predictor, based on which the effect of the environmental parameters can be calculated. The ANN’s one-way sensitivity analysis was performed with the use of “PaD” algorithm. The most contributing parameters were found to
be the WT and SRP parameters. The synergistic effect of the parameters was calculated with the use of the “PaD2” algorithm, a paired sensitivity analysis method. The paired parameters interactions revealed that the SRP-WT combination had the higher relative contribution to the simulated Chl-\(a\). These results showed that the SRP is the lake’s key nutrient and that the lake algal production is highly affected by climatic change.

The DIN had no major relative contribution. However the role of DIN must not be underestimated in the restoration attempts, since the ANN results showed that the synergistic effect of SRP-DIN has a clear association pattern with the algal production. Taking all these into consideration, it is indicated that the management measures taken for lake restoration must be updated and new restoration practices should be applied alongside the existing ones. This ANN model combined with the “PaD” and “PaD2” algorithms has become a successful management guiding tool revealing trends between Chl-\(a\) production and the environmental variables.

5.5. References

Atoui, A.; Hafez, H.; Slim, K. Occurrence of toxic cyanobacterial blooms for the first time in Lake Karaoun, Lebanon. *Water Environ.* 2013, **27**, 42–49.

Cui, Y.; Zhu, G.; Li, H.; Luo, L.; Cheng, X.; Jin, Y.; Trolle, D. Modelling the response of phytoplankton to reduced external nutrient load in a subtropical Chinese reservoir using DYRESM-CAEDYM. *Lake Reserv. Manag.* 2016, **32**, 146–157.

Dedecker, A.; Goethals, P.; D’Heygere, T.; Gevrey, M.; Lek, S.; De Pauw, N. Application of Artificial Neural Network models to analyse the relationships between *Gammarus Pulex* L. (Crustacea, Amphipoda) and river characteristics. *Environ. Monit. Assess.* 2005, **111**, 223–241.
Demuth, H.; Beale, M.; Hagan, M. *Neural Network Toolbox User’s Guide*; MathWorks, Inc.: Natick, MA, USA, 2007.

Dimopoulos, I.; Bourret, P.; Lek, S. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Process. Lett.* **1995**, *2*, 1–4.

Dodds, W.K.; Bouska, W.W.; Eitzmann, J.L.; Pilger, T.J.; Pitts, K.L. Eutrophication of U.S. freshwaters: Analysis of potential economic damages. *Environ. Sci. Technol.* **2009**, *43*, 12–19.

Dyhrman, S.T. Molecular approaches to diagnosing nutritional physiology in harmful algae: Implications for studying the effects of eutrophication. *Harmful Algae* **2008**, *8*, 167–174.

Ferrante, M.; Oliveri Conti, G.; Fiore, M.; Rapisarda, V. Harmful algal blooms in the Mediterranean Sea: Effects on human health. *Euromediterr. Biomed. J.* **2013**, *8*, 25–34.

Ferreira, J.G.; Andersen, J.H.; Borja, A.; Bricker, S.B.; Camp, J.; Cardoso da Silva, M.; Garcés, E.; Heiskanen, A.S.; Humborg, C.; Ignatiades, L.; et al. Overview of eutrophication indicators to assess environmental status within the European Marine Strategy Framework Directive. *Estuar. Coast. Shelf Sci.* **2011**, *93*, 117–131.

Ferreira, J.G.; Bricker, S.B.; Simas, T.C. Application and sensitivity testing of a eutrophication assessment method on coastal systems in the United States and European Union. *J. Environ. Manag.* **2007**, *82*, 433–445.

Feuchtmayr, H.; Moran, R.; Hatton, K.; Connor, L.; Heyes, T.; Moss, B.; Harvey, I.; Atkinson, D. Global warming and eutrophication: Effects on water chemistry and autotrophic communities in experimental hypertrophic shallow lake mesocosms. *J. Appl. Ecol.* **2009**, *46*, 713–723.
Gevrey, M.; Dimopoulos, I.; Lek, S. Two-way interaction of input variables in the sensitivity analysis of neural network models. *Ecol. Model.* **2006**, *195*, 43–50.

Gevrey, M.; Dimopoulos, I.; Sovan, L. Review and comparison of methods to study the contribution of variables in Artificial Neural Networks models. *Ecol. Model.* **2003**, *160*, 249–264.

Gkelis, S.; Papadimitriou, T.; Zaoutsos, N.; Leonardos, I. Anthropogenic and climatic-induced change favors toxic cyanobacterial blooms: Evidence from monitoring a highly eutrophic, urban Mediterranean Lake. *Harmful Algae* **2014**, *39*, 322–333.

Gubelit, Y.I.; Berezina, N.A. The causes and consequences of algal blooms: The *Cladophora glomerata* bloom and the Neva estuary (Eastern Baltic Sea). *Mar. Pollut. Bull.* **2010**, *61*, 183–188.

Guneralp, B.; Barlas, Y. Dynamic modelling of a shallow freshwater lake for ecological and economic sustainability. *Ecol. Model.* **2003**, *167*, 115–138.

Ha, M.H.; Pflugmacher, S. Phytotoxic effects of the cyanobacterial neurotoxin anatoxin-α: Morphological, physiological and biochemical responses in aquatic macrophyte, *Ceratophyllum demersum*. *Toxicon* **2013**, *70*, 1–8.

Hoeger, S.J.; Hitzfeld, B.C.; Dietrich, D.R. Occurrence and elimination of cyanobacterial toxins in drinking water treatment plants. *Toxicol. Appl. Pharmacol.* **2005**, *203*, 231–242.

Hu, Y.H.; Hwang, J.N. *Handbook of Neural Network Signal Processing*; CRC Press: Boca Raton, FL, USA, 2002.

Jeppesen, E.; Kronvang, B.; Olesen, J.E.; Audet, J.; Sondergaard, M.; Hoffmann, C.C.; Andersen, H.E.; Lauridsen, T.L.; Liboriussen, L.; Larsen, S.E.; et al. Climatic change effects on nitrogen loading from cultivated catchments in Europe: Implications for
nitrogen retention, ecological state of lakes and adaptation. *Hydrobiologia* **2011**, 663, 1–21.

Kagalou, I.; Papastergiadou, E.; Leonardos, I. Long term changes in the eutrophication process in a shallow Mediterranean Lake ecosystem of W. Greece: Response after the reduction of external load. *J. Environ. Manag.* **2008**, 87, 497–506.

Kagalou, I.; Papastergiadou, E.; Tsimarakis, G.; Petridis, D. Evaluation of the trophic state of Lake Pamvotis Greece, a shallow urban lake. *Hydrobiologia* **2003**, 506, 745–752.

Karul, C.; Soyupak, S.; Çilesiz, A.F.; Akbay, N.; Germen, E. Case studies on the use of neural networks in eutrophication modeling. *Ecol. Model.* **2000**, 134, 145–152.

Karul, C.; Soyupak, S.; Yurteri, C. Neural network models as a management tool in lakes. *Hydrobiologia* **1999**, 408, 139–144.

Kazemi Yazdi, S.; Scholz, M. Assessing the storm water detention systems treating road runoff with an Artificial Neural Network predicting fecal indicator organisms. *Water Air Soil Pollut.* **2010**, 206, 35–47.

Koreiviene, J.; Anne, O.; Kasperoviciene, J.; Burskyte, V. Cyanotoxin management and human health risk mitigation in recreational waters. *Environ. Monit. Assess.* **2014**, 186, 4443–4459.

Kuo, J.T.; Hsieh, M.H.; Lung, W.S.; She, N. Using Artificial Neural Network for reservoir eutrophication prediction. *Ecol. Model.* **2007**, 200, 171–177.

LaBounty, J.F. Secchi transparency of Boulder Basin, Lake Mead, Arizona-Nevada: 1990–2007. *Lake Reserv. Manag.* **2008**, 24, 207–218.

Lek, S.; Guégan, J.F. Artificial Neural Networks as a tool in ecological modelling, an introduction. *Ecol. Model.* **1999**, 120, 65–73.
Liu, Y.; Chen, W.; Li, D.; Huang, Z.; Shen, Y.; Liu, Y. Cyanobacteria-/cyanotoxin-contaminations and eutrophication status before Wuxi Drinking Water Crisis in Lake Taihu, China. *J. Environ. Sci.* **2011**, **23**, 575–581.

Maier, H.R.; Dandy, G.C.; Burch, M.D. Use of Artificial Neural Networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. *Ecol. Model.* **1998**, **105**, 257–272.

Mateus, M.; Almeida, C.; Brito, D.; Neves, R. From eutrophic to mesotrophic: Modelling watershed management scenarios to change the trophic status of a reservoir. *Int. J. Environ. Res. Public Health* **2014**, **11**, 3015–3031.

Mooij, W.M.; De Senerpont Domis, L.N.; Janse, J.H. Linking species- and ecosystem-level impacts of climate change in lakes with a complex and a minimal model. *Ecol. Model.* **2009**, **220**, 3011–3020.

Moss, B.; Kosten, S.; Meerhoff, M.; Battarbee, R.; Jeppesen, E.; Mazzeo, N.; Havens, K.; Lacerot, G.; Liu, Z.; De Meester, L.; et al. Allied attack: Climate change and eutrophication. *Inland Waters* **2011**, **1**, 101–105.

Muttil, N.; Chau, K.W. Machine-learning paradigms for selecting ecologically significant input variables. *Eng. Appl. Artif. Intell.* **2007**, **20**, 735–744.

Özesmi, S.L.; Tan, C.O.; Özesmi, U. Methodological issues in building, training, and testing artificial neural networks in ecological applications. *Ecol. Model.* **2006**, **195**, 83–93.

Paerl, H.; Huisman, J. Blooms like it hot. *Science* **2008**, **320**, 57–58.

Paerl, H.W. Assessing and managing nutrient-enhanced eutrophication in estuarine and coastal waters: Interactive effects of human and climatic perturbations. *Ecol. Eng.* **2006**, **26**, 40–54.

Paerl, H.W.; Xu, H.; McCarthy, M.J.; Zhu, G.; Qin, B.; Li, Y.; Gardner, W.S. Controlling harmful cyanobacterial blooms in a hyper-eutrophic lake (Lake Taihu, China): The
need for a dual nutrient (N & P) management strategy. *Water Res.* **2011**, **45**, 1973–1983.

Papastergiadou, E.; Kagalou, I.; Stefanidis, K.; Retalis, A.; Leonardos, I. Effects of anthropogenic Influences on the trophic state, land uses and aquatic vegetation in a shallow Mediterranean Lake: Implications for restoration. *Water Resour. Manag.* **2010**, **24**, 415–435.

Papatheodorou, G.; Demopoulou, G.; Lambrakis, N. A long-term study of temporal hydrochemical data in a shallow lake using multivariate statistical techniques. *Ecol. Model.* **2006**, **193**, 759–776.

Sagehashi, M.; Sakoda, A.; Suzuki, M. A mathematical model of a shallow and eutrophic lake (the Keszthely Basin, Lake Balaton) and simulation of restorative manipulations. *Water Res.* **2001**, **35**, 1675–1686.

Sagrario, M.A.G.; Jeppesen, E.; Gomà, J.; Søndergaard, M.; Jensen, J.P.; Lauridsen, T.; Landkildehus, F. Does high nitrogen loading prevent clear-water conditions in shallow lakes at moderately high phosphorus concentrations? *Freshw. Biol.* **2005**, **50**, 27–41.

Salami Shahid, E.; Ehteshami, M. Application of Artificial Neural Networks to estimating DO and salinity in San Joaquin River Basin. *Desalination Water Treat.* **2016**, **57**, 4888–4897.

Scardi, M.; Harding, L.W., Jr. Developing an empirical model of phytoplankton primary production: A neural network case study. *Ecol. Model.* **1999**, **120**, 213–223.

Sener, E.; Terzi, O.; Sener, S.; Kucukkara, R. Modeling of water temperature based on GIS and ANN techniques: Case study of Lake Egirdir (Turkey). *Ekoloji* **2012**, **21**, 44–52.

Stefanidis, K.; Papastergiadou, E.S. Aquatic vegetation and related abiotic environment in a shallow urban lake of Greece. *Belg. J. Bot.* **2007**, **140**, 25–38.
Tota-Maharaj, K.; Scholz, M. Artificial Neural Network simulation of combined permeable pavement and earth energy systems treating storm water. *J. Environ. Eng.* **2012**, *138*, 499–509.

Trolle, D.; Staehr, P.A.; Davidson, T.A.; Bjerring, R.; Lauridsen, T.L.; Søndergaard, M.; Jeppesen, E. Seasonal dynamics of CO\(_2\) flux across the surface of shallow temperate lakes. *Ecosystems* **2012**, *15*, 336–347.

van Maanen, B.; Coco, G.; Bryan, K.R.; Ruessink, B.G. The use of Artificial Neural Networks to analyze and predict alongshore sediment transport. *Nonlinear Process. Geophys.* **2010**, *17*, 395–404.

Wang, L.; Kisi, O.; Zounemat-Kermani, M.; Hu, B.; Gong, W. Modeling and Comparison of hourly photosynthetically active radiation in different ecosystems. *Renew. Sustain. Energy Rev.* **2016**, *56*, 436–453.

Xu, F.L.; Tao, S.; Dawson, R.W.; Li, B.G. A GIS-based method of lake eutrophication assessment. *Ecol. Model.* **2001**, *144*, 231–244.

Ye, C.; Shen, Z.; Zhang, T.; Fan, M.; Lei, Y.; Zhang, J. Long term effect of nutrients and temperature increase on algal growth in Lake Taihu, China. *J. Environ. Sci.* **2011**, *23*, 222–227.

Zhang, P.; Hong, B.; He, L.; Cheng, F.; Zhao, P.; Wei, C.; Liu, Y. Temporal and spatial simulation of atmospheric pollutant PM2.5 changes and risk assessment on population exposure to pollution using optimization algorithms of the back propagation-Artificial Neural Network model and GIS. *Int. J. Environ. Res. Public Health* **2015**, *12*, 12171–12195.

Zounemat-Kermani, M. Principal Component Analysis (PCA) for estimating chlorophyll concentration using forward and generalized regression neural networks. *Appl. Artif. Intell.* **2014**, *28*, 16–29.
6. Application 2: Evaluating the contributing environmental parameters associated with eutrophication in a shallow lake by applying Artificial Neural Networks Techniques.

6.1. Introduction

Environmental pollution is a serious problem affecting the modern world, therefore several studies were carried out in order to examine the consequences of pollution on the environment and human health like the one conducted by Grigoropoulos et al. (2008). The most examined pollutants used to be the micro elements but in recent studies the toxic metals started to be taken into consideration as well, because of their negative impact on the environment and human health (Aydin et al., 2015). Special emphasis is given on the pollution of lakes because they are the major providers of freshwater world-wide, covering several human needs like drinking and irrigation. It is estimated that lakes contain about 90% of the liquid freshwater on earth’s surface (Mishra and Garg, 2011). Surface water eutrophication is considered the most widespread problem to water environment quality around the world (Wu et al., 2011). The causes of eutrophication are closely related to nutrients in-puts, mainly phosphorus and nitrogen, into the water column through anthropogenic activities such as sewage discharge and agriculture (Aydin et al, 2015; Hou et al., 2013; Wang et al., 2013). In many cases eutrophication might cause a Harmful Algal Bloom (HAB) with many unpleasant consequences for a lake, like fish kills or scum formation (Atou et al., 2013). A possible way to prevent HABs is through monitoring. Monitoring of water resources is a way to protect
them from pollution and to prevent environmental contamination from pollutants like fertilizers and pesticides (Yagmur et al., 2014).

The negative impact of eutrophication on water quality led to the development of several models trying to simulate it. Artificial Neural Net-works (ANNs) ability to process problems with non linear and complex data, even if the data are imprecise and noisy (Lek and Guegan, 1999), has made the ANNs ideal for lake eutrophication modelling. Another advantage of ANNs regarding Chl-$a$ forecasting is that they can satisfy the modelling objectives by using only routine monitoring data (Li et al., 2010). In our case a time lagged ANN is created for Chl-$a$ simulation. Some examples of ANN models used for fore-casting Chl-$a$ with weekly time-lag are described by Muttil and Chau (2007); Palani et al. (2008); Park et al. (2015) while ANN models with monthly time-lag are described by Belgrano et al. (2001); Li et al. (2007).

Figure 6.1. Map of the study area combined with a map of Greece, where Ioannina city is located northwestern. The urban area of Ioannina city is observed in the west part of the lake.
The aim of this modelling study was the development of a one month ahead forecasting time series model, which would simulate the Chl-α levels. The created ANN model has the ability to predict possible HABs, giving the local authorities the chance to intervene and prevent the HAB event. However, the primary purpose of this modelling study is to evaluate the impact of each environmental parameter associated with the eutrophication phenomenon. With the usage of sensitivity analysis techniques the effect of the environmental parameters can be measured and so to have an insightful of the eutrophication process. By that way the created ANN model is given the ability to simulate different management scenarios for its parameters, especially for the nutrients, and therefore to produce a guideline for several restoration techniques that could be applied for water quality improvement.

6.2. Materials and Methods

6.2.1. Monitoring Area and Monitored Data.

Pamvotis Lake is located next to the city of Ioannina, in northwestern Greece (Figure 6.1). More details about Lake Pamvotis are given in Chapter 5.

The used data were from a monitoring study that was carried out for the chronological period between June 2004 and August 2005. The data were collected on a monthly basis from eleven monitoring stations within the lake. The measured parameters were soluble reactive phosphorus (SRP), dissolved inorganic nitrogen (DIN) species (nitrite, nitrate, ammonium), carbonate (CO₃), bicarbonate (HCO₃), pH, electric conductivity (EC), water temperature (WT), dissolved oxygen (DO), secchi disk (SD) and chlorophyll-α (Chl-α).
6.2.2. Artificial Neural Networks Methodologies.

A very popular ANN is the multi-layer feedforward network. Feedforward networks are usually divided into three layers and each layer is consisted by neurons. The first layer is the input layer, one intermediate hidden layer and the output layer that gives the final result. Each neuron in a layer is connected with all the neurons in the next layer with a synaptic weight. Aggregation is performed on the weighted inputs of every neuron from the previous layer and an output value is yielded through a transfer function (Recknagel et al., 1997). The output value of the jth neuron ($o_j$) is given by the equations:

\[ o_j = f(u_j) \]  
(Eq. 6.1)

and

\[ u_j = \sum w_{ij} x_i + z_j \]  
(Eq. 6.2)

where $f$ is the transfer function; $x_i$ is the input from ith neuron from the previous layer; $w_{ij}$ is the synaptic weight that connects $x_i$ with the jth neuron and $z_j$ a bias term. The output of each neuron is computed and propagated to the next layer and the network output is compared with the given output. The learning procedure is repeated for several times with the use of a training algorithm and each time the synaptic weights are adjusted until they minimize an error function, usually taken the mean square difference between the predicted and the given output (Sengorur et al., 2006; Salamishahid and Ehteshami, 2016). The methods of cross-validation and regularization are two approaches widely used to avoid over-fitting phenomenon (Dedecker et al., 2005). Over-fitting occurs when the ANN has big error for new data simulations making the model unable to generalize well new situations.

The topology of a 3 layer ANN can be represented as L1-H1-L2, where L1 is the number of neurons in the input layer, H1 the number of neurons in the hidden layer and L2 the number of neurons in the output layer. To ensure that there aren't too many neurons into the
hidden layer and so the ANN starts over-fitting, some rules of thumb are proposed but no assured methods exist until now (Li et al., 2007). In our case the number of the hidden layer neurons is according to the rules proposed by Maier et al. (1998):

\[ N^H \leq 2N^I + 1 \]  
(Eq. 6.3)

and

\[ N^H \leq 2N^{TR}/(N^I + 1) \]  
(Eq. 6.4)

where \( N^H \) is the number of hidden layer neurons, \( N^I \) the number of inputs and \( N^{TR} \) the number of training samples. The maximum \( N^H \) must be the minimum number calculated by those two equations.

Several methodologies have been proposed for evaluating an ANN model's sensitivity for its input parameters and finding their contribution. In this modelling study three different sensitivity techniques are applied. The first one is the ‘Perturb’ method. The ‘Perturb’ method is computing the perturbation effect of the input variables regarding the output variable (Dedecker et al., 2005). The effect that a small change of an input variable has on the ANN's output is examined and the input variables are classified by an order of importance (Gevrey et al., 2003). The ‘Weights’ method is related with the interpretation of the connections weights from the input layer to the hidden layer (Muttil and Chau, 2007). The contribution of each input parameter to the ANN's output is measured with the relative parameter importance (I) equation, which is given by Lee et al. (2003) as bellow:

\[ I = \frac{\sum_{n=1}^{nT} \sum_{j=1}^{nH} |(wij)|}{\sum_{k=1}^{nT} (\sum_{n=1}^{nT} \sum_{j=1}^{nH} |(wij)|_k)_{nT}} \]  
(Eq. 6.5)

where \( nT \) is the number of time lags; \( nH \) is the number of hidden neurons and \( nV \) is the number of input parameters.
The ‘PaD’ method or the Partial Derivatives method was firstly described by Dimopoulos et al. (Dimopoulos et al., 1995). This method is following two steps. First the partial derivatives of the output for small changes of each input are computed and afterwards the classification of the relative contributions of each input variable ance (Gevrey et al., 2003). The sensitivity of the ANN’s output for the input variable \(x_i\), is symbolized as \(SSD_i\) and according to Dimopoulos et al. (1995) is the sum of the squared partial derivatives obtained per input variable and has the following form:

\[
SSD = \sum_{j=1}^{N} (d_{ji})^2
\]

(Eq. 6.6)

where \(N\) is the observations number of the data set; \(d_{ji}\) the partial derivatives of the output \(y_j\) with respect to input \(x_j\). The equation for \(d_{ji}\) is given bellow:

\[
d_{ji} = S_j \sum_{h=1}^{n_h} w_{ih} \ast I_h \ast (1 - I_h) \ast w_{ho}
\]

(Eq. 6.7)

where \(n_h\) the neurons number of the hidden layer; \(I_h\) the output of the \(h\)-th hidden neuron for the \(j\)-th input; \(w_{ih}\) the weight connecting the \(i\)-th input neuron and the \(h\)-th hidden neuron; \(w_{ho}\) the weight connecting the output and the \(h\)-th hidden neuron; \(S_j\) is the derivative of the output \(y_j\) with respect to input \(x_j\).

**Table 6.1.** ANN’s performance for the training, test and whole set, based on the RMSE, root mean square error; MAE, absolute mean error and \(r\), correlation coefficient.

| Model Performance | Training set | Test set | Whole set |
|-------------------|--------------|----------|-----------|
| MAE               | 12.77        | 15.75    | 13.36     |
| RMSE              | 25.87        | 28.46    | 26.41     |
| \(r\)             | 0.940        | 0.959    | 0.944     |
6.2.3. ANN Model Development.

The ANN's training algorithm was the Levenberg–Marquardt (LM) algorithm, because the LM algorithm has the fastest convergence among the variations of back-propagation algorithm when the data set is up to few hundreds parameters (Hagan et al., 1996). The method of regularization was used to avoid model's over-fitting. The data set consists of 150 samples was divided into two subsets: a) the training set, which contains 80% of the measured data and b) the test set, which contains 20% of the measured data. The model's performance was evaluated for the test set. The ANN's performance was evaluated with the use of correlation coefficient ($r$), Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) (Table 1). The simulations were computed with the use of MatLab software. In order to eliminate the magnitude differences among the variables, the data set was normalized before the training and then the ANN output values were unnormalized for the Chl-a predictions (Hou et al., 2007).

In order to reduce the model's dimensionality and therefore computational complexity Principal Components Analysis (PCA) is applied to the data set. Following the methodology described by Akratos et al. (2008); Zounemat-Kermani (2014) the ANN is simulated with and without the candidate parameters combinations. As proposed by Sener et al. (2012) the ANN was simulated several times with different number of neurons in the hidden layer for finding the optimal topology. The chosen combination of layers number and neurons number must be the one that gives the closer simulated output to the target (Salami shahid and Ehteshami, 2016).

Among the tested topologies the best topology was the 8-10-1, with $r=0.959$ for the test set and $r=0.94$ for the learning set. The best combination of input parameters was SRP$_{(t-1)}$, DIN$_{(t-1)}$, pH$_{(t-1)}$, EC$_{(t-1)}$, WT$_{(t-1)}$, DO$_{(t-1)}$, SD$_{(t-1)}$ and Chl-a$_{(t-1)}$. The model's output is Chl-a, where $t$ is the time variable measured in months. A synopsis of the model’s description is
given in Table 6.2. The parameter Chl-a(t-1) was included as an input since Chl-a levels are influenced by their previous values (Palani et al., 2008) and algal production has nonstationary behavior (Hou et al., 2007). The introduction of the time lagged parameter Chl-a(t-1) as a tactic used in many ANN studies simulating Chl-a like the ones of Muttil and Chau (2007); Palani et al. (2008); Wu et al. (2014). The PCA revealed that the carbonate parameters have small contribution to the total variation. Additionally, ANN’s performance was higher when the carbonates weren’t included as an input parameter (Table 6.3). Therefore, the carbonate parameters were excluded as input parameters, since they have no real impact on Chl-a simulations.

Table 6.2. Synoptic presentation of the ANN model, presenting the models input parameters, output parameter, learning function and topology.

| ANN Model | Inputs* | Output | Learning function | Topology |
|-----------|---------|--------|-------------------|----------|
| SRP(t-1), DIN(t-1), pH(t-1), EC(t-1), WT(t-1), DO(t-1), SD(t-1) and Chl-a(t-1) | Chl-a(t) | LM algorithm | 8-10-1 |

*Where t parameter corresponds to time in months

Table 6.3. ANN performance evaluation for the two inputs sets for the RMSE, root mean square error; MAE, absolute mean error and r, correlation coefficient.

| Model Inputs | r   | MSE | MAE |
|--------------|-----|-----|-----|
| Set 1: SRP(t-1), DIN(t-1), pH(t-1), EC(t-1), WT(t-1), DO(t-1), SD(t-1) and Chl-a(t-1) | 0.959 | 28.46 | 15.75 |
| Set 2: SRP(t-1), DIN(t-1), pH(t-1), EC(t-1), WT(t-1), DO(t-1), SD(t-1) and Chl-a(t-1), CO3(t-1), HCO(t-1) | 0.923 | 32.62 | 19.35 |
Figure 6.2. ANN model simulated data plotted against the Chl-a measured data for the monitoring period July 2004 - August 2005.

Figure 6.3. Parameter contribution with the use of 'Perturb' method.

Figure 6.4. Parameter contribution with the use of 'Weights' method.

Figure 6.5. Parameter contribution with the use of 'PaD' method.

6.3. Results and Discussion

The graphical representation of the real against the simulated Chl-a values (Figure 6.2) shows that the ANN simulated well the Chl-a with most of the simulated values being very similar with the measured ones, so the created ANN model can be characterized as reliable. In
some cases, the measured data that correspond to extreme Chl-\(a\) values are not matching very well with the simulated data, increasing by that way the ANN’s error. However, the ANN model managed to characterize these extreme data correctly, as eutrophic or hyper-trophic. The ‘Perturb’ method was applied in order to examine the effect of each input parameter on the model’s output. Each input parameter was increased for \(\pm 10\%\) and the change of the model output was calculated. The results of the ‘Perturb’ method (Figure 6.3) are showing the simulated output tendency for reduction or increase when a specific input parameter is fluctuated. It is simulated that an increase of a parameter’s arithmetical value is associated with an increase of Chl-\(a\) levels, except for the case of EC parameter that has a reduction of Chl-\(a\) levels. This is associated with the fact that the EC is negatively correlated with Chl-\(a\) concentrations and the lowest EC values are observed during a HAB (Atoui et al., 2013). The WT according to the ‘Perturb’ method is the most influential parameter, a finding verified by a study of Paerl and Huisman (2008).

The application of the ‘Weights’ method illustrated the relative contribution of each parameter to the simulated Chl-\(a\) output (Figure 6.4). The input parameter Chl-\(a(t-1)\) was the most influential. Some other ANN modelling studies of eutrophication e.g., Muttill and Chau (2007); Scardi and Harding (1999), verify our findings since they found the input parameter Chl-\(a(t-1)\) as the most important. The SRP followed second confirming the high correlation of SRP with Chl-\(a\) (jeppesen et al., 2006). The WT, SD, DO, EC, pH and DIN were next regarding relative parameter importance.

The results of the ‘PaD’ method (Figure 6.5) discriminated the input parameters into two groups. The first group has parameters with high relative contribution. In this group the DO parameter has the higher relative contribution, followed by Chl- \(a(t-1)\), SRP, SD and pH parameters. The second group has input parameters with almost zero contribution and is consisted of DIN, WT and EC. This discrimination between minor and major contributing
parameters with the use of the ‘PaD’ method is also observed by Mouton et al. (2008) and was attributed to the high sensitivity of the ‘PaD’ method.

The results obtained by the sensitivity methods led to useful conclusions, highlighting the need for these results to be incorporated into a unique arithmetical value for each input parameter that will be used for the evaluation of the environmental parameters. An importance index for each input parameter is created, that assigns each parameter an importance order value from the highest to the lowest contributing parameter. For every sensitivity method the importance order index is calculated for each input parameter. The mean importance order value is calculated for the input parameters based on the results obtained by the three sensitivity methods (Table 6.4). Therefore, the mean importance order index is the combined interpretation of the results given by the three sensitivity methods. By that way the computational differences found for the input parameters when are calculated with a different sensitivity method are balanced. These differences are explained by the different calculation procedures followed by each sensitivity technique (Dedecker et al., 2005).

The mean importance index found that the Chl-\(\alpha(t-1)\) input parameter has the most significant role for the eutrophication process. The DO input parameter was calculated to have the second most important role. This is attributed to the fact that the DO parameter has a strong connection with the input parameter Chl-\(\alpha(t-1)\), since the lake is supplied with oxygen by algal photosynthesis (Misra et al., 2011). The major contribution of SRP to the eutrophication process is verified. The SRP is ranked third according to the mean importance index. The WT parameter follows next to the fourth place. The ‘PaD’ method calculated the WT as the less important parameter; however the importance of WT is emphasized by the results of ‘Weights’ and ‘Perturb’ methods. The SD and pH parameters were in the fifth place. The linkage between eutrophication and high pH values especially during summer is
mentioned by Trolle et al. (2012). The SD parameter it is inversely correlated with Chl-\(\alpha\) concentrations (Canfield et al., 2016). The DIN and EC parameters were the least contributing parameters according to the mean importance index. Low importance of the DIN and EC is found by all the three sensitivity methods. The result for DIN is confirmed by Sondergaard et al. (2011) study, where total nitrogen was correlated significantly with Chl-\(\alpha\) but with a smaller correlation than for phosphorus.

6.4. Conclusions

ANN models can successfully simulate multi-parameter modelling scenarios because of their ability to process non-linear relationships between the dependent and independent variables. In our case the created ANN managed to produce simulated outputs highly correlated with the measured data, making our model a reliable predictor. The application of the three sensitivity techniques allowed us a thorough look into the contributing parameters and their role regarding eutrophication. The introduction of the mean importance order index gave as the advantage to balance the computational borne differences resulted after applying the three different sensitivity methods.

The necessity for a modelling study before any restoration attempt is obvious, because the decision makers can identify the correct restoration technique for the lake. A lake model that combines parameters like planktivorous fish, zooplankton and nutrients is proposed for a future modelling study regarding lake Pamvotis. Unfortunately, no such data set is available in our case, so only the nutrient reduction effect can be examined. The SRP has a leading contributing role regarding eutrophication, in contrast with DIN that was found to have the smallest impact on the model. Therefore, the SRP can be characterized as the key nutrient of the lake and special emphasis must be given on SRP reduction. Nevertheless, the DIN
contribution to eutrophication must not be underestimated since based on ‘Perturb’ method results it is observed that an increase of DIN causes an increase of Chl-α values. A further combined reduction of SRP and DIN is proposed, since the nutrients are associated mainly with anthropogenic activities indicating the need for extra restoration measures for the improvement of lake's trophic status.

**Table 6.4.** Parameter importance order index ranking the time lagged parameters from the most to the least contributing. For the ‘PaD’ method; ‘Perturb’ method; ‘Weights’ method the analog index is assigned. The mean importance index is the mean index value resulted by the three mentioned methods.

| Input parameter | ‘PaD’ importance index | ‘Perturb’ importance index | ‘Weights’ importance index | Mean importance index |
|-----------------|------------------------|----------------------------|----------------------------|-----------------------|
| SRP             | 3                      | 5                          | 2                          | 3                     |
| DIN             | 6                      | 7                          | 8                          | 6                     |
| pH              | 5                      | 2                          | 7                          | 5                     |
| EC              | 7                      | 8                          | 6                          | 6                     |
| WT              | 7                      | 1                          | 3                          | 4                     |
| DO              | 1                      | 3                          | 5                          | 2                     |
| SD              | 4                      | 6                          | 4                          | 5                     |
| Chl-α           | 2                      | 4                          | 1                          | 1                     |
6.5. References

Akratos, C.; Papaspyros, J.; Tsihrintzis, V. An artificial neural network model and design equations for BOD and COD removal prediction in horizontal subsurface flow constructed wetlands. *Chemical Engineering Journal* 2008, 143, 96-110.

Atoui, A.; Hafez, H.; Slim, K. Occurrence of toxic cyanobacterial blooms for the first time in Lake Karaoun, Lebanon. *Water and Environment* 2013, 27, 42-49.

Aydin, S.; Yagmur, B.; Coban, H.; Simsek, H. The nutritional conditions and some heavy metal contents of the vineyards in a semi-arid area. *Fresenius Environmental Bulletin* 2015, 24, 1-6.

Belgrano, A.; Malmgren, B. A.; Lindahl, O. Application of artificial neural networks (ANN) to primary production time-series data. *Journal of Plankton Research* 2001, 23(6), 651-658.

Canfield, D.; Bachmann, R.; Stephens, D.; Hoyer, M.; Bacon, L., et al. Monitoring by citizen scientists demonstrates water clarity of Maine (USA) lakes is stable, not declining, due to cultural eutrophication. *Inland Waters* 2016, 6, 11-27.

Dedecker, A.; Goethals, P.; D'Heygere, T.; Gevrey, M.; Lek, S.; De Pauw, N. Application of Artificial Neural Network models to analyse the relationships between Gammarus Pulex L. (Crustacea, Amphipoda) and river characteristics. *Environmental Monitoring and Assessment* 2005, 111, 223-241.

Dimopoulos, I.; Bourret, P.; Lek, S. Use of some sensitivity criteria for choosing networks with good generalization ability. *Neural Process. Lett.* 1995, 2, 1–4.

Gevrey, M.; Dimopoulos, I.; Sovan, L. Review and comparison of methods to study the contribution of variables in artificial neural networks models. *Ecological Modelling* 2003, 160, 249-264.
Grigoropoulos, K.N.; Nastos, P.N.; Ferentinos, G.; Gialouris, A.; Vassiliou, T., et al. Spatial Distribution of PM1 and Sinus Arrhythmias in Athens, Greece. *Fresenius Environmental Bulletin* **2008**, *17*, 1426-1431.

Hagan, M.T.; Demuth, H.B.; Beale, M. Neural Network Design. PWS Publishing: Boston, USA, 1996.

Hou, D.; He, J.; Lü, C.; Sun, Y.; Zhang, F. et al. Effects of Environmental Factors on Nutrients Release at Sediment-Water Interface and Assessment of Trophic Status for a Typical Shallow Lake, Northwest China. *The Scientific World Journal* **2013**, 10.1155/2013/716342, 1-16.

Hou, G.; Li, H.; Recknagel, F.; Song, L. Using the moving window incorporated neural network to forecast the population behavior of Nostocales spp. in the river Darling, Australia. *Fresenius Environmental Bulletin* **2007**, *16*, 304-309.

Jeppesen, E.; Pekcan-Hekim, Z.; Lauridsen, T. L.; Søndergaard, M.; Jensen, J.P. Habitat distribution of fish in late summer: changes along a nutrient gradient in Danish lakes. *Ecology of Freshwater Fish* **2006**, *15*, 180–190.

Lee, J.H.W.; Huang, Y.; Dickman, M.; Jayawardena, A.W. Neural network modelling of coastal algal blooms. *Ecological Modelling* **2003**, *159*, 179-201.

Lek, S.; Guégan, J.F. Artificial neural networks as a tool in ecological modelling, an introduction. *Ecological Modelling* **1999**, *120*, 65-73.

Li, H.; Hou, G.; Dakui, F.; Xiao, B.; Song, L., et al. Prediction and elucidation of the population dynamics of Microcystis spp. in Lake Dianchi (China) by means of artificial neural networks. *Ecological Informatics* **2007**, *2*, 184–192.
Li, H.; Hou, G.; Song, G.; Liu, Y. Modelling Algal Blooms in Lake Dianchi, China using Neural Networks. *Fresenius Environmental Bulletin* **2007**, *16*, 798-803.

Li, X.; Nour, M.H.; Smith, D.W.; Prepas, E. Neural networks modelling of nitrogen export: model development and application to unmonitored boreal forest watersheds. *Environmental Technology* **2010**, *31*, 495-510.

Maier, H. R.; Dandy, G. C.; Burch, M.D. Use of artificial neural networks for modelling cyanobacteria *Anabaena* spp. in the River Murray, South Australia. *Ecological Modelling* **1998**, *105*, 257-272.

Mishra, A.K.; Garg, N. Analysis of Trophic State Index of Nainital Lake from Landsat – 7 ETM Data. *J. Indian Soc. Remote Sens.* **2011**, *39*, 463–471.

Misra, A.K.; Chandra, P.; Raghavendra, V. Modeling the depletion of dissolved oxygen in a lake due to algal bloom: Effect of time delay. *Advances in Water Resources* **2011**, *34*, 1232–1238.

Mouton, A.; Dedecker, A.; Lek, S.; Goethals, P. Selecting Variables for Habitat Suitability of *Asellus* (Crustacea, Isopoda) by Applying Input Variable Contribution Methods to Artificial Neural Networks Models. *Environ Model Assess* **2010**, *15*, 65-79.

Muttil, N.; Chau, K.W. Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Applications of Artificial Intelligence* **2007**, *20*, 735–744.

Paerl, H.; Huisman, J. Blooms Like It Hot. *Science* **2008**, *320*, 57-58.

Palani, S.; Liong, S.Y.; Tkalich, P. An ANN application for water quality forecasting. *Marine Pollution Bulletin* **2008**, *56*, 1586–1597.
Park, Y.; Cho, K.H.; Park, J.; Cha, S.M.; Kim, J.H. Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater and estuaries reservoirs, Korea. *Science of the Total Environment* **2015**, *502*, 31-41.

Recknagel, F.; French, M.; Harkonen, P.; Yabunaka, K.I. Artificial neural network approach for modelling and prediction of algal blooms. *Ecological Modelling* **1997**, *96*, 11-28.

Salami Shahid, E.; Ehteshami, M. Application of artificial neural networks to estimating DO and salinity in San Joaquin River basin. *Desalination and water Treatment* **2016**, *57*, 4888-4897.

Scardi, M.; Harding Jr., L.W. Developing an empirical model of phytoplankton primary production: a neural network case study. *Ecological Modelling* **1999**, *120*, 213–223.

Sener, E.; Terzi, O.; Sener, S.; Kucukkara, R. Modeling of Water Temperature Based on GIS and ANN Techniques: Case Study of Lake Egirdir (Turkey). *Ekoloji* **2012**, *21*, 44-52.

Sengorur, B.; Dogan, E.; Koklu, R.; Samandar, A. Dissolved oxygen estimation using artificial neural network for water quality control. *Fresenius Environmental Bulletin* **2006**, *15*, 1064-1067.

Søndergaard, M.; Larsen, S. E.; Jørgensen, T.; Jeppesen, E. Using chlorophyll a and cyanobacteria in the ecological classification of lakes. *Ecological Indicators* **2011**, *11*, 1403–1412.

Trolle, D.; Staehr, P. A.; Davidson, T. A.; Bjerring, R.; Lauridsen, T. L. et al. Seasonal Dynamics of CO2 Flux Across the Surface of Shallow Temperate Lakes. *Ecosystems* **2012**, *15*, 336–347.
Wang, L.; Liu, L.; Zheng, B. Eutrophication development and its key regulating factors in a water supply reservoir in North China. *Journal of Environmental Sciences* **2013**, *25*, 962–970.

Wu, H.; Zhang, J.; Li, P.; Zhang, J.; Xie, H. et al. Nutrient removal in constructed microcosm wetlands for treating polluted river water in northern China. *Ecological Engineering* **2011**, *37*, 560-568.

Wu, N.; Huang, J.; Schmalz, B.; Fohrer, N. Modeling daily chlorophyll a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches. *Limnology* **2014**, *15*, 47-56.

Yagmur, B.; Aydin, S.; Okur, B.; Coban, H.; Simsek, H. Quality parameters of vineyard irrigation water in a semi-arid region: the plain of Alasehir, Turkey. *Fresenius Environmental Bulletin* **2014**, *23*, 2650-2655.

Zounemat-Kermani, M. Principal Component Analysis (PCA) for estimating chlorophyll concentration using forward and generalized regression neural networks. *Applied Artificial Intelligence* **2014**, *28*, 16-29.
7. Application 3: Assessment of the Eutrophication-Related Environmental Parameters in Two Mediterranean Lakes by Integrating Statistical Techniques and Self-Organizing Maps.

7.1. Introduction

Freshwater quality has declined in the last decades throughout Europe, due to various environmental issues related to anthropogenic activities. Eutrophication is considered as one of the most important environmental problems that affects freshwater, coastal, and marine ecosystems worldwide (Yang et al., 2008). Freshwater lakes are major providers of water for several purposes, such as water supply, drinking, irrigation, and so forth. Eutrophication has a negative impact on water quality, with ecological and socioeconomic consequences.

Eutrophication triggers various physical and chemical changes in the aquatic environment that may cause the blooming of certain harmful-toxin-producing algae (cyanophyta), which are known to create health issues for organisms living in the lakes as well as humans (Ignatiades and Gotsis-Skretas, 2012). Because of the adverse effects of eutrophication, the necessity for mitigating eutrophic phenomena and recovering water quality has become a priority for environmental scientists (Yang et al., 2008). The effect of eutrophication on public health is so serious that cyanobacteria have been characterized as potential key hazardous pollutants by the European Water Framework Directive (2000) (2000/60/EC) (Hoeger et al., 2004). Water sports, such as swimming, in eutrophic lakes and the consumption of seafood and drinking water
contaminated with cyanotoxins are the main reasons for human illness related to cyanotoxins. It is reported that about 60,000 intoxication incidents, with an overall mortality rate of 1.5%, take place per year globally because of algal toxins (Ferrenet et al., 2013).

Artificial neural networks (ANNs) are considered to be a computational modelling tool that is widely used in solving many complex real-world problems (Maier and Dandy, 2001; Odabas, 2014). In recent years, ANNs have become a desirable tool that is applied in many scientific topics (Moustris et al., 2011), such as air pollution, precipitation, water quality, and classification of rainfall prediction (Nastos et al., 2013). The ability of ANNs to learn from the known data without being affected by nonlinearity and their classification potential qualify them as being superior to other traditional statistical tools (Simsek, 2016). Modelling studies that used both ANNs and multiple linear regression (MLR) methods found that ANNs produced better modelling results than MLR in most cases (Odabas, 2014). The main weakness of MLR models as compared with ANN models is that they are based on a linear relationship between input and target variables (Simsek, 2016).

ANNs are divided into two main categories: ANNs with supervised learning and ANNs with unsupervised learning. The most popular supervised ANN is the multilayer perceptron with a backpropagation algorithm, while the most popular unsupervised ANN is the Kohonen self-organizing map (SOM) (Park et al., 2003). According to Peeters et al. (2007), the SOM has recently started to be used in exploratory data analysis, e.g., Tsai et al. (2017), Ejarque-Gonzalez & Butturini (2014), and it can help with summarizing available data and extracting useful information. The SOM can deal with the phenomenon of nonlinearity, handle noisy data, and be updated easily (Lee and Scholz, 2006a); because of this potential, the SOM algorithm is considered as a powerful tool in exploratory data analysis and the clustering of multivariate data.
sets (Peeters et al., 2007). The importance of the SOM for water quality management is significant, as the SOM has the potential to analyze multidimensional ecological data and simplify them into visual information that is helpful to understanding the ecological process (Lee and Scholz, 2006b).

Application of the SOM algorithm is widely used in the environmental sciences, and especially in studies examining water quality. For example, Recknagel et al. (2006) used SOMs to evaluate the seasonality effect over two adjacent lakes in response to eutrophication control. A SOM was used by Oh et al. (2007) to cluster the phytoplankton communities from a reservoir. In the study of Cheng et al. (2012), the SOM was used to classify fish communities in shallow lakes, based on several biotic and abiotic factors such as water depth, transparency, and dissolved oxygen. The multi-relationships between fish species and river water quality parameters were examined with the use of a SOM by Tsai et al. (2017). In their study, Park et al. (2003) used a SOM to classify 23 different water types, such as streams, lakes, rivers, canals, and ponds. The SOM classified the sampling sites into five clusters based on environmental parameters, and related them with species richness. In another modelling study, Tota-Maharaj and Scholz (2013) applied the SOM to simulate microbial data from the effluent of pavement systems used to treat stormwater runoff, with a good accuracy, having a minimum correlation of $R=0.751$ between the real and predicted data, and using as the model’s inputs parameters that are not expensive to measure; whereas the measurement of microbial pathogens concentrations is a time-consuming and expensive procedure. These examples demonstrate the applicability of the SOM algorithm not only in ecological research, but also in water management studies.

The objective of this study was twofold. First, we assessed the interactions of the environmental parameters related to the algal productivity of two transboundary Greek Lakes-
Megali Prespa (or Great Prespa) and Mikri Prespa (or Small Prespa)—by applying a combination of a PCA, a cluster analysis, and a SOM algorithm. Second, we compared the results among the different techniques to evaluate the effectiveness of SOMs in relation to more “traditional” tools such as PCA and cluster analysis. The importance of this work also lies within the fact that there is no other published work using a SOM to study the water quality parameters related to the eutrophication of Greek lakes, to the knowledge of the authors. Therefore, this study could set the basis for the modelling with SOMs of additional Greek lakes that are impacted by the effects of eutrophication. Additionally, the results of this study can be useful for exploring the mechanisms associated with algal productivity in the studied area.

7.2. Materials and Methods

7.2.1. Study Area and Data Collection

The transboundary Prespa area is a geographically remote area located in northwestern Greece. It is an area of great ecological importance, and has been declared a national park (in 1974), a Ramsar wetland of international importance (in 1987), an important bird area (in 1983), and a Natura 2000 site (Pyrovetsi, 1989). Lakes Megali Prespa, at 849m above sea level (asl) (Panagiotopoulos et al., 2013), and Mikri Prespa, at 853.5m asl (Koussouris et al., 1989), are surrounded by mountains and located in the transboundary Prespa area shared by Greece, Albania, and the Former Yugoslavian Republic of Macedonia (FYROM) (Figure 7.1).

The climate of the area is characterized as sub-Mediterranean with continental influences, with frequent snowfall in the winter and summer rain drops (Panagiotopoulos et al., 2013). The transboundary catchment’s area covers about 1300 km² (Panagiotopoulos et al., 2013), and the
permanent population of the Greek part is estimated at about 1500 residents (Pyrovetsi, 1989). Megali Prespa is a large, deep lake, covering 254 km², with a 14 m mean and 48 m maximum water depth (Cvetkoska et al., 2016; Kagalou and Leonardos, 2009). Mikri Prespa is a shallow lake, with approximately 48 km² of surface area, and a maximum water depth of 8 m (Loffler et al., 1998; Albrecht et al., 2012; Stefanidis and Papastergiadou, 2010) and mean depth of 4.1 m (Koussouris et al., 1989; Albrecht et al., 2012; Stefanidis and Papastergiadou, 2010). The two lakes formed a single lake in the past, but nowadays are distinct and connected through an artificial channel (Panagiotopoulos et al., 2013). Lake Mikri Prespa is supplied with water only seasonally through surface runoff, mainly from small rivers (Koussouris et al., 1989), and overflows into Lake Megali Prespa (Cvetkoska et al., 2016; Leng et al., 2013).

**Figure 7.1.** Map of the studied transboundary Lakes Megali and Mikri Prespa in northwestern Greece, with the sampling sites marked in red.

The inflow from Lake Mikri Prespa is about 9%, while the rest of the water inflow into Megali Prespa is from small streams (56%) and direct precipitation (35%) (Aufgebauer et al., 2012). Lake Megali Prespa has no surface outflow, but is connected through karstic channels to the neighbouring Lake Ohrid (Panagiotopoulos et al., 2013).
The trophic status of Lake Mikri Prespa is characterized as eutrophic, and prolonged cyanobacterial blooms occur, which may start in spring and persist until December, favoured by the warm climate (Vardaka et al., 2005). Lake Megali Prespa is considered mesotrophic Albrecht et al., 2012; Leng et al., 2013), with summer bottom anoxia and an average total phosphorus concentration of 31 mg·m⁻³ (Leng et al., 2013).

Table 7.1. Statistical description of the seasonally measured environmental parameters for Lakes Megali Prespa and Mikri Prespa.

| Variable                      | Mikri Prespa (n=79) | Megali Prespa (n=26) |
|-------------------------------|---------------------|----------------------|
|                               | Mean | Minimum | Maximum | Mean | Minimum | Maximum |
| pH                            | 7.93 | 7.00    | 8.60     | 8.32 | 8        | 8.90    |
| Dissolved oxygen (mg/L)       | 9.96 | 4.50    | 18.00    | 9.75 | 5.90    | 13.00   |
| Electrical conductivity (μS/cm)| 281.94   | 310.00 | 310.00 | 221.61 | 227.00 |
| Secchi depth (m)              | 0.92 | 0.40    | 2.00     | 3.35 | 1.00    | 6.00    |
| Water depth (m)               | 1.39 | 0.70    | 6.33     | 1.50 | 1.50    | 12.00   |
| Water temperature (°C)        | 19.20 | 12.60 | 26.1     | 18.54 | 13.80 | 24.00   |
| Total phosphorus (μg/L)       | 123.47 | 463.00 | 249.00 | 77.43 | 21.50 | 249.10 |
| Dissolved inorganic nitrogen (mg/L) | 319.07 | 2486.00 | 808.30 | 249.00 | 249.07 | 79.50   |
| Chlorophyll-a (mg/m³)         | 10.76 | 1.10    | 42.70    | 4.01 | 0.40    | 14.50   |

Water samples were collected from a total of fifteen sampling sites in Lake Mikri Prespa and four sites in Lake Megali Prespa (Figure 7.1). Samplings were carried out on a seasonal basis during spring, summer, and autumn, for the monitoring period of 2006–2008 (Stefanidis and Papastergiadou, 2010). The measured environmental parameters were: pH, surface dissolved oxygen (DO), electrical conductivity (EC), Secchi disk (SD) depth, water depth, surface water
temperature (WT), total phosphorus (TP), dissolved inorganic nitrogen (DIN), and chlorophyll-a (Chl-a).

The parameters were measured at several sampling sites within the Greek part of the lakes for each monitoring season. Most of the samples from Lake Megali Prespa were collected from the littoral zone of the lake near the Psarades village, located south of Lake Megali Prespa, and no samples from the pelagic zone were included. The statistical description of the data is given in Table 7.1. More details regarding the monitoring process and the measurement of environmental parameters are given by Stefanidis & Papastergiadou (2010).

7.2.2. Statistical Methods and Theoretical Background

Principal component analysis (PCA) is a multivariate data analysis method used to reveal patterns in large data sets (Hardle and Simar, 2003). PCA is a mathematical dimension reduction procedure, used to reduce the number of variables of a data set to a smaller number of variables, without information loss of the initial data set (Birks, 2012). PCA transforms the data into a new set of variables (or coordinates), called principal components (PCs) (Hardle and Simar, 2003). The use of principal components instead of the initial variables is a more reliable way to represent relationships, because the system’s noise is reduced (Mueller and Grunsky, 2016). In this modelling study, the environmental parameters were normalized prior to any calculations, and all the samples that contained missing values were removed from the data set. The parameters were standardized by subtracting the sample mean from each observation and then dividing by the sample standard deviation. The simulation results for the statistical methods were
carried out using the MatLab software. Only the PCs that explain at least 10% of the variance are considered for this study, as the other PCs contributed very little.

Cluster analysis is a technique that classifies objects into different groups based on their characteristics (Reimann et al., 2008). The aim of cluster analysis is to find groups (clusters) with homogeneous properties out of heterogeneous large samples, with high internal homogeneity within clusters and high heterogeneity between clusters (Hardle and Simar, 2003). High levels of similarity among objects are indicated by a small value in a distance matrix and large values in proximity or similarity matrices (Gore, 2000). Cluster methods are divided into four categories: hierarchical methods, partitioning methods, overlapping cluster procedures, and ordination techniques (Hardle and Simar, 2003).

Euclidean distance and Ward agglomerative methods were used for the cluster analysis. The clusters number can be calculated with the use of a cutoff line for the dendrogram, based on the Sneath’s index of cluster significance (Sneath and Sokal, 1973). The less restrictive significance criterion of Sneath’s index (2/3 of $D_{\text{max}}$) and the strict significance criterion of Sneath’s index (1/3 of $D_{\text{max}}$) may be used, where $D_{\text{max}}$ is the maximum of the distance measure $D$.

### 7.2.3. Self Organizing Map Theory

The Kohonen SOM is an unsupervised ANN. The SOM has the ability to learn without being given the associated output values for the corresponding input data, and the desired output is not known a priori. The goal of the learning process is to classify the input data according to their similarity (Peeters et al., 2007). SOMs are a very practical tool for data visualization; also, SOMs can be used for prediction and correlation analysis, mostly with visual representation (Rub et al, 2009). One of the various applications of the SOM algorithm is the finding of statistically
significant dependencies among the variables in a multidimensional data sample, where two highly correlated variables produce two similar component planes (Barreto-Sanz and Perez-Uribe, 2007). A SOM projects high-dimensional data into a low-dimension space (Kohonen, 2001). This is usually a two-dimensional space, so the neurons are arranged in two dimensions (see Figure 7.2), because the visual summary of the output is more understandable (Aguilera et al., 2001). The sample data can be clustered either as manually determined by a U-matrix or can be automated by a clustering algorithm implemented in the SOM, and usually the hierarchical clustering algorithm is applied (Vesanto and Alhoniemi, 2000; Choi et al., 2014).

![Figure 7.2. Projection of n-dimensional data into a two-dimensional (XY) space (output layer) with the use of a SOM; where x1, x2,...,xn are the input variables, n is the input variable’s number, and wij is the synaptic weight that is connecting the i input variable with the j node.

Figure 7.2. Projection of n-dimensional data into a two-dimensional (XY) space (output layer) with the use of a SOM; where x1, x2,...,xn are the input variables, n is the input variable’s number, and wij is the synaptic weight that is connecting the i input variable with the j node.

The SOM is consisted of an input layer and an output layer that are connected with computational weights (Park et al., 2006; An et al., 2016). The output layer consists of neurons that are arranged in a hexagonal or rectangular grid and are fully interconnected (Peeters et al., 2007). The input patterns, usually after normalization, are imported through the neurons in the input layer. The SOM algorithm can be summarized into the following steps (Choi et al., 2014):

1. Weight vector initialization with random values.
2. Use of a distance measure, usually the Euclidean distance, to find the best-matching unit (BMU).

3. Movement closer to the input vector by updating the weight vector of the BMU and the neighboring neurons.

The Euclidean distance \( D_i \) mentioned above is described by the following equation, and calculates the distance measure between the input vector and the \( i \) weight vector (An et al., 2016; Zhang et al., 2008):

\[
D_i = \sqrt{\sum_{j=1}^{R} (p_{ij} - w_{ij})^2} \quad ; \quad i = 1,2, ..., S \quad \text{(Eq. 7.1)}
\]

where \( S \) is the number of output neurons, \( R \) is the dimension of the input vectors, \( p_{ij} \) represents the \( j \) element of the input vector, and \( w_{ij} \) symbolizes the \( j \) element of the \( i \) weight vector. The term BMU is defined, according to Lee & Scholz (2006b), as the neuron with the weight vector closest to the input variable \( x \), as given by the equation:

\[
|x - m_c| = \min(|x - m_i|) \quad \text{(Eq. 7.2)}
\]

where \( |x| \) symbolizes the distance measure, \( x \) the input vector, \( m \) the weight vector, and \( c \) the subscription of the weight vector for the winning neuron. A more detailed description of the SOM algorithm can be found in the studies of Lee & Scholz (2006b), Aguilera et al. (2001), Park et al. (2006) and An et al. (2016).
All the input parameters are transformed before presenting them to the SOM neural network, with the use of log transformation. The optimum map size of the SOM can be calculated based on the relationship:

$$M \approx 5\sqrt{n}$$  \hspace{1cm} (Eq. 7.3)

where $n$ is the data sample number and $M$ is the number of neurons (An et al., 2016; Zhang et al., 2008). The SOM Toolbox for MatLab (Vesanto et al., 2000) was used for data simulations.

7.3. Results

7.3.1. PCA and Cluster Analysis Results

The first four principal components (PCs) of the PCA analysis satisfy the criterion of the proportion of variance accounted for (Hardle and Simar, 2003), and together they summarize 74.17% of the total variance. A Pareto chart is created, where the percentage of variance explained by each PC is presented (Figure 7.3). At the same time, the criterion of eigenvalue-one is met, as suggested by An et al. (2016), by retaining the first four PCs (principal components).
**Figure 7.3.** Pareto chart representing the percentage of variance explained by the principal components (PCs) of the PCA analysis.

A descriptive table for the first four PCs and their characteristics is created (Table 7.2).

**Table 7.2.** Principal components (PCs) percentage of variance, eigenvalues, and loadings for the examined environmental parameters.

| Variable | PC1       | PC2       | PC3       | PC4       |
|----------|-----------|-----------|-----------|-----------|
| pH       | −0.1609   | **0.6158**| 0.1653    | −0.0569   |
| DO       | −0.0350   | −0.0237   | **0.7513**| −0.1349   |
| EC       | **0.4908**| −0.0198   | −0.0993   | −0.0736   |
| SD       | −**0.5300**| 0.0511    | −0.2700   | −0.1709   |
| Depth    | −**0.5171**| 0.1026    | −0.2527   | −0.1701   |
| WT       | 0.2051    | **0.6144**| −0.1386   | 0.0279    |
| TP       | 0.1613    | −0.2545   | **−0.4452**| 0.1072    |
| DIN      | 0.1248    | −0.1500   | −0.0275   | **−0.9302**|
| Chl–a    | **0.3230**| **0.3766**| −0.2085   | −0.1936   |
| **Eigenvalue** | 2.94 | 1.55 | 1.18 | 1.01 |
| **Variance explained (%)** | 32.68 | 17.29 | 13.08 | 11.12 |
| **Cumulative variance (%)** | 32.68 | 49.97 | 63.05 | 74.17 |

DO: dissolved oxygen; EC: electrical conductivity; SD: Secchi disk depth; WT: surface water temperature; TP: total phosphorus; DIN: dissolved inorganic nitrogen; Chl–a: chlorophyll-a.

The significant correlation coefficients are shown with bold characters.

PC1 explains 32.68% of the total variance, and has strong negative loadings for SD and depth, and strong positive loadings for EC and phytoplankton Chl-a. PC1 seems to describe the
relationships between SD, depth, EC and Chl-a. PC2 explains 17.29% of the total variance, and has strong positive loadings for pH, WT, and Chl-a. PC2 describes the relationship between the Chl-a and WT parameters. PC3 explains 13.08% of the total variance, and has a strong positive loading for DO and a strong negative loading for TP. PC4 explains 11.12% of the total variance, and has a strong negative loading for DIN. The bivariate plots for the PCs (Figure 7.4) reveal differences among the samples from Lake Megali Prespa, symbolized as L1 (n1=26), and the samples from Lake Mikri Prespa, symbolized as L2 (n2=89). The bivariate plots of PC1 versus the rest of the PCs show a clear separation between the samples from the two lakes.

![Bivariate plots](image)

**Figure 7.4.** Bivariate plots between four different principal components (PCs) of Lake Megali Prespa (L1) and Lake Mikri Prespa (L2).

The cluster analysis grouped the samples with respect to the water quality parameters. The cluster analysis divided the data into three clusters, where Ward’s linkage method and Euclidean distance measurements were used. The dendrogram constructed by the means of cluster analysis, shown in Figure 7.5, illustrates the arrangement of the clusters.
Figure 7.5. A dendrogram of similarity for the data samples from Lake MegaliPrespa (L1) and Lake MikriPrespa (L2). Based on Sneath’s criterion (red horizontal line), three clusters are formed (separated by the red vertical solid lines). Dmax: maximum distance.

Cluster 3 corresponds to samples from Lake Megali Prespa, and cluster 1 and cluster 2 correspond to samples from Lake Mikri Prespa. A clear distinction is observed for the data from the two lakes, where all L1 samples are illustrated at the right side (branch) of the dendrogram and all L2 samples are illustrated at the left side (branch) of the dendrogram. The Sneath’s index of cluster significance calculated the cluster number with the use of a cutoff line for the dendrogram; the less restrictive criterion of Sneath’s index of cluster significance, equal to 2/3 of $D_{max}$, was used.

7.3.2. SOM Algorithm Results

A SOM with $10 \times 5$ neurons was created, based on the rule given by Equation (7.3). The components planes (CPs) are visualized in Figure 7.6. For the clustering of the SOM prototypes, the K-means algorithm method and the hierarchical algorithm method were used and evaluated in order to find the most appropriate clustering method for this modelling study.
Figure 7.6. Visualization of the SOM’s component planes (CPs) for each environmental parameter, where the colorbars indicate the mapping of the data values.

The optimal number of clusters minimizes the Davies–Bouldin index when the SOM is implemented by the K-means algorithm (An et al., 2016; Wang et al., 2014). In our case, the optimal cluster number was five (Figure 7.7). The L1 samples (Megali Prespa) correspond to cluster 1 and cluster 2, while the L2 samples (Mikri Prespa) correspond to cluster 3, cluster 4, and cluster 5. The SOM gave a clear classification of the samples from the two lakes.
Figure 7.7.(a) Clustering of the SOM based on the K-means algorithm. (b) Hits histogram of SOM analysis representing the density of SOM hits. The green represents data from Lake Mikri Prespa, the red represents data from Lake Megali Prespa, and the empty nodes are associated with the absence of data samples.

The hits histogram distinguished the Megali Prespa data from the Mikri Prespa data. The difference between the data from Megali Prespa and data from Mikri Prespa is primarily associated with the water transparency (or SD). It is also reversely associated with the EC.

The hierarchical cluster analysis performed on the SOM prototypes calculated three district groups for the investigated lakes. Ward’s linkage method and Euclidean distance measurement were used, and a dendrogram was derived (Figure 7.8). The existence of three clusters was computed with the use of the less restrictive significance of Sneath’s index (2/3 of $D_{max}$). Based on the dendrogram, the L1 samples (Megali Prespa) are located on the right side (branch) and belong to cluster 3, while the L2 (Mikri Prespa) samples are located on the left side (branch) and belong to cluster 1 and cluster 2.

Figure 7.8. A dendrogram of similarity for the SOM’s prototype nodes. Based on Sneath’s less restrictive criterion (red horizontal line), three clusters are formed (separated by the red vertical solid lines).
As expected, the sampling sites from the shallower Lake Mikri Prespa were associated with higher Chl-a values, as shallow lakes are more prone to nutrient-mixing eutrophication that might benefit algal production (Stefanidis and Papastergiadou, 2012). The TP values also increased as water depth decreased. It is documented that strong positive relationships exist between TP loadings and algal biomass (Hadjisolomou et al., 2017). No clear conclusions are extracted for the DO parameter. The pH parameter seems to have a negative relationship with the Chl-a, and generally the pH parameter is associated with eutrophication (Kagalou and Leonardos, 2009; Atoui et al., 2013). Besides the visualization of the results with the CPs, the SOM clusters can also provide useful information regarding the water quality parameter interactions.

For the clustering of the SOM’s prototypes, two different clustering methods were used. The K-means algorithm method (see Table 7.3) and the hierarchical algorithm method (see Table 7.4) were applied. The mean values and standard deviation (SD) for each SOM cluster were calculated and are presented in Tables 3 and 4.

**Table 7.3.** Statistical description of the SOM’s clusters found using the K-means algorithm method, based on mean values and standard deviation (SD).

| Variable | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 | Cluster 5 |
|----------|-----------|-----------|-----------|-----------|-----------|
|          | Mean     | SD        | Mean     | SD        | Mean     | SD        | Mean     | SD        | Mean     | SD        |
| pH       | 8.4      | 0.22      | 8.18     | 0.15      | 7.91     | 0.35      | 7.76     | 0.33      | 8.06     | 0.39      |
| DO       | 10.19    | 2.05      | 9        | 1.34      | 11.54    | 2.74      | 9.82     | 2.65      | 9.55     | 1.99      |
| EC       | 220.06   | 3.65      | 229.45   | 19.89     | 276.13   | 15.31     | 265.14   | 16.95     | 290.21   | 12.24     |
| SD       | 3.23     | 1.78      | 3.33     | 1.78      | 0.98     | 0.34      | 0.96     | 0.28      | 0.85     | 0.33      |
| WT       | 20.38    | 3.6       | 16.25    | 1.73      | 15.73    | 0.35      | 15.54    | 3.74      | 23.36    | 2.19      |
| TP       | 77.38    | 41.46     | 73.61    | 57.9      | 145.5    | 2.74      | 112.89   | 43.98     | 122.6    | 80.34     |
| DIN      | 331.76   | 232.62    | 135.12   | 36.23     | 481.37   | 15.31     | 101.75   | 36.62     | 195.54   | 83.21     |
| Chl-a    | 5.64     | 3.7       | 1.9      | 0.72      | 6.37     | 0.34      | 7.22     | 3.36      | 15.4     | 9.74      |
The K-means algorithm is considered a most suitable clustering method for the result implementation of this modelling study, as it calculated five clusters instead of the three that the hierarchical algorithm found. The existence of five clusters instead of three allows the more detailed examination of the environmental parameter interactions. For example, regarding the Lake Megali Prespa-associated clusters, the role of WT and therefore the seasonality effect is easily observed in Table 7.3, but no clear conclusions can be derived from Table 7.4.

**Table 7.4.** Statistical description of the SOM’s clusters found using the hierarchical algorithm method, based on mean values and standard deviation (SD).

| Variable | Cluster 1 | Cluster 2 | Cluster 3 |
|----------|-----------|-----------|-----------|
|          | Mean | SD | Mean | SD | Mean | SD |
| pH       | 8.07 | 0.37 | 7.77 | 0.33 | 8.30 | 0.22 |
| DO       | 9.62 | 2.37 | 10.78 | 2.54 | 9.66 | 1.84 |
| EC       | 289  | 11.89 | 267.69 | 17.17 | 224.24 | 14.04 |
| SD       | 0.89 | 0.38 | 0.94 | 0.22 | 3.27 | 1.75 |
| WT       | 22.82 | 2.66 | 14.68 | 2.16 | 18.30 | 3.55 |
| TP       | 126.64 | 81.25 | 124.46 | 55.43 | 75.71 | 48.45 |
| DIN      | 183.44 | 82.82 | 316.05 | 326.88 | 244.36 | 199.01 |
| Chl-a    | 13.41 | 9.96 | 7.35 | 3.2 | 3.98 | 3.34 |

**4. Discussion**

Unraveling and investigating trophic function mechanisms is a task of major importance, particularly for developing sustainable management and restoration plans for lakes with poor water quality. There is a large number of monitoring studies that have pointed out that nutrient enrichment has been the main cause of eutrophication, for example Atoui et al. (2013); Kagalou
et al. (2008). PCA and cluster analysis are statistical methods commonly used in such studies that often provide satisfactory results and useful conclusions about the main parameters associated with eutrophication processes.

Conversely, SOMs models are considered to be more advanced modelling tools used in water quality modelling studies (Wang et al., 2014). Modelling studies examining lake water quality by combining a SOM model and statistical methodologies are approaching more comprehensively the behaviour of the limnological parameters. Wang et al. (2014) and An et al. (2016) displayed valuable results with statistical methodologies, however the findings of the SOM model were much more practical and specific. The prevalence of the SOM method against PCA is related to the fact that the relationships among environmental variables are nonlinear, while the PCA is based on linear principles (Muttil and Chau, 2007). An important advantage of the SOM compared with the PCA method is based on its visualization abilities, provided by the SOM’s component planes. With the use of the component planes, the distribution of the component values is represented and direct visual examination is provided in order to allow correlations between several component planes that can be investigated simultaneously (Cinar and Merdun, 2009).

Another advantage of the SOM method compared with the classic statistical techniques is related to the dimension reduction provided by the SOM by projecting multidimensional data into a two-dimensional space. The dendrogram created by applying the hierarchical clustering method on the standardized raw data is usually overcrowded, making its implementation difficult. In contrast, the dendrogram resulting from the SOM prototypes is easily implemented. This is because the SOM neural network abridges the samples of the initial data set to a smaller
set of prototypes, providing the possibility to construct a simplified dendrogram (Goncalves et al., 2008).

Representation by the CPs can be considered as an efficient and practical visualization tool for extracting useful conclusions regarding the trophic function of the two lakes. The L1 (Megali Prespa) data are correlated with high SD and low EC values, thus providing a clear separation from L2 (Mikri Prespa) data. The parameter SD seems to have a crucial role for clustering the data of the studied lakes. Meanwhile, the SD parameter has a negative correlation with the EC, since shallower lakes are often associated with higher EC. The high WT are linked with high Chl-a values, mainly for samples from Lake Mikri Prespa. Despite the rather small number of samples collected from the littoral zone of the Greek part of Lake Megali Prespa, the SOM managed to separate the data from the two lakes, and produced results comparable to those of the other statistical tools. The environmental parameters associated with algal production were visualized by the SOM component planes. Through the CPs, the visualization of all variables enables them to be examined simultaneously, and direct associations between the variables for specific value ranges of each examined variable can be obtained. In addition, PCA and cluster analysis do not always capture the hidden information that is provided by the data set. In contrast, the SOM can detect hidden patterns and recognize specific features of the data set which are often different in the short-term assessment (Astel et al., 2007). The SOM revealed hidden patterns for each lake after careful examination of the CPs (Figure 7.6). Our results showed that the modelled environmental parameters, which are associated with algal production, followed a pattern influenced by the seasonality effect for each lake data set. Specifically, the K-means algorithm clustering results for Lake Mikri Prespa showed that the data of cluster 5 had higher values for the Chl-a and WT parameters than the other two clusters, suggesting the
seasonality effect over Lake Mikri Prespa and the elevated algal production during hot months. The data that are associated with cluster 3 and cluster 4 had moderate values for Chl-a and WT, suggesting such conditions of these parameters during spring and autumn. However, no clear patterning is observed for the rest of the parameters. A deeper examination of the revealed patterns shows that for Lake Mikri Prespa, Chl-a presents its maximum value when the WT is also maximum; while at the same time, the TP presents its minimum and the DIN is relatively very low. Regarding Lake Megali Prespa, the minimum Chl-a value is associated with the maximum SD value, elevated WT value, and relative low TP and DIN values. In contrast, the maximum Chl-a value is associated with the maximums of the pH and WT parameter values, elevated DIN, and a very low TP value. These two clusters suggest an almost reversed patterning between them, where Lake Megali Prespa is affected by the seasonality to a lesser extent than the shallow Lake Mikri Prespa.

Additionally, the SOM more reliably visualizes nonlinear and heterogeneous data than the PCA method into a two-dimensional space (Brosse et al., 2001). The above observations highlight the SOM superiority compared with the PCA, since the SOM captures well the complex nonlinear mechanisms associated with algal production. However, the PCA did not manage to relate the parameter interactions well enough, limited by its linear nature as a method. For example, the nonlinear relationship between the Chl-a and TP parameters is well presented by the SOM through its CPs as was discussed for each lake data set, and specific information regarding the parameter interactions can be extracted based on the mapping of the data values. In contrast, the PCA (see Figure 7.4) provides only generalized information for a tendency regarding these parameter interactions. Because of this, the SOM is considered to be an innovating method compared with PCA and cluster analysis. Based on the SOM results, high
nutrient concentrations were associated with elevated phytoplankton Chl-a values. High water temperature (WT) values were also correlated with the high Chl-a values. The role of water temperature as a key environmental parameter for controlling phytoplankton biomass in Lake Mikri Prespa was highlighted in a study by Tryfon & Moustaka-Gouni (1997), where it was shown that an increase in biomass of cyanophytes occurred at temperatures greater than 16°C. The SD depth was also found to be a critical parameter based on PCA, cluster analysis, and SOM. Both the cluster analysis method and the SOM method managed to distinguish between the samples from Lakes Mikri and Megali Prespa, mainly because of the SD depth variation. However, the SOM with the use of K-means algorithm distinguished the water quality parameters into five diverse clusters, providing a deeper examination of the parameter interactions, in contrast with the cluster analysis method, which gave only three clusters. The results presented in Table 7.3 showed that cluster 2 samples are characterized by lower Chl-a and lower WT values than cluster 1, reflecting the seasonal variations of algal productivity of Lake Megali Prespa. Additionally, cluster 3 and cluster 4, which are associated with Mikri Prespa, show noticeably different mean WT values compared with cluster 5. Cluster 5, in particular, seems to be associated with the hot months and the highest temperatures, where Chl-a has the highest mean value. On the other hand, cluster 3 and cluster 4 have lower WT and Chl-a mean values, probably representing conditions typical of the spring and autumn seasons.

The depth parameter, which is strongly correlated with SD depth, was not included as an input variable for the SOM model, because the aim of this modelling study was to focus on the mechanisms that are affecting algal production, and not the morphological characteristics of the lakes. In order to avoid bias in the results, it was decided that the depth parameter would not be included as an input parameter.
It is documented that low water transparency (SD depth) is associated with increased algal production (Jeppesen et al., 2007; Scavia et al., 2014), and Chl-a has an analogous relationship with the SD depth (Hadjisolomou et al., 2016). In a study by Zacharias et al. (2002) that examined the biological and chemical characteristics of the major Greek lakes, it was stated that the SD depth is high in deep lakes and low in shallow lakes. The SOM model that managed to separate the two lakes matched the high SD depth with samples from Lake Megali Prespa and low SD depth values with samples from Lake Mikri Prespa. As expected, the observation of the CPs associated the data samples of Lake Mikri Prespa with low SD depth values and increased Chl-a levels, and the data samples of Lake Megali Prespa with high SD depth values and low Chl-a levels.

The role of the SD depth has been given great attention in many studies, in order to discriminate possible environmental factors influencing the water quality between shallow and deep lakes. For example, in a study by Stefanidis et al. (2016), two geomorphologically different lakes were examined, and it was found that the SD depth was a good indicator of the water quality deterioration and enhancement of eutrophication. The above monitoring examples, combined with the fact that the nutrient loading threshold for obtaining clear water conditions differs among climate zones (Jeppesen et al., 2017), lead to the conclusion that modelling limnological data sets is a very case-sensitive task; while the mechanisms controlling limnological parameters are complex and their interactions are hard to examine. Therefore, the application of a SOM is ideal in studying water quality parameters and their interactions in lakes.

The increasing eutrophication that has been reported for the studied lakes (Albrecht et al., 2012) is mainly related to human activities. The last decades’ increased land use activities in the catchment’s area, such as agriculture, the encroachment of wetland areas, and the expansion of
the irrigation system, have had a negative impact on the wetland biodiversity, enhancing the eutrophication processes (Pyrovetsi, 1989). This water quality deterioration was demonstrated by Loffler et al. (1998), where a remarkable decrease of Secchi depth in Lake Megali Prespa, from 7.2 to 10m in the 1950s to 3.2m in the 1990s, was noted. The SOM managed to simulate the algal production process for the two lakes and identify the environmental parameter interactions. Therefore, SOM algorithms can be used as a tool that can aid management authorities in developing a guideline regarding the restoration measures that should be applied. Besides the water transparency that was one of the key factors associated with water quality, the water temperature was significantly correlated with phytoplankton Chl-a. Based on this finding, it would be interesting to further explore the effects of the anticipated rising of temperature, along with intensification of water level changes on the water quality and the processes of eutrophication.

In our case, the SOM separated the two sampling areas and verified the good water quality status of Lake Megali Prespa. Regarding the shallow Lake Mikri Prespa, which is suffering the effects of eutrophication, the role of water transparency and water temperature should be considered when prioritizing management measures. Proposed management options should mitigate the excessive use of fertilizers, but should also consider other pressures, such as water abstraction, in order to reduce nutrient loadings and provide a relatively stable water level regime.

7.5. Conclusions

This modelling study presented different methodologies in order to examine the water quality of the investigated lakes. The results derived from the classic statistical methodologies
were compared with the SOM neural network findings. The results from PCA, cluster analysis, and the SOM method agreed about the trophic function properties of the lakes, but the SOM results were more specific, and allowed direct associations between the water quality variables. The SOM neural network has a basic advantage derived from its visualization abilities. Based on our findings, the SOM neural network can be described as an innovating modelling tool that can be used autonomously or in parallel with the rest of the traditional modelling methods for successfully evaluating freshwater quality.

7.6. References

Aguilera, P.A.; GarridoFrenich, A.; Torres, J.A.; Castro, H.; Martinez Vidal, J.L.; Canton, M. Application of the Kohonen Neural Network in Coastal Water Management: Methodological Development for the Assessment and Prediction of Water Quality. *Water Resour.* **2001**, *17*, 4053–4062.

Albrecht, C.; Hauffe, T.; Schreiber, K.; Wilke, T. Mollusc biodiversity in a European ancient lake system: Lakes Prespa and MikriPrespa in the Balkans. *Hydrobiologia* **2012**, *682*, 47–59.

An, Y.; Zou, Z.; Li, R. Descriptive Characteristics of Surface Water Quality in Hong Kong by a Self-Organising Map. *Int. J. Environ. Res. Public Health* **2016**, *13*, 115, doi:10.3390/ijerph13010115.

Astel, A.; Tsakovski, S.; Barbieri, P.; Simeonov, V. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Res.* **2007**, *41*, 4566–4578.

Atoui, A.; Hafez, H.; Slim, K. Occurrence of toxic cyanobacterial blooms for the first time in Lake Karaoun, Lebanon. *Water Environ. J.* **2013**, *27*, 42–49.

Aufgebauer, A.; Panagiotopoulos, K.; Wagner, B.; Schaebitz, F.; Viehberg, F.; Vogel, H.; Zanchetta, G.; Sulpizio, R.; Leng, M.J.; Damaschke, M. Climate and environmental change in the Balkans over
the 17 ka recorded in sediments from Lake Prespa (Albania/F.Y.R. of Macedonia/Greece). *Quat. Int.* **2012**, *274*, 122–135.

Barreto-Sanz, M.A.; Perez-Uribe, A. Improving the correlation hunting in a large quantity of SOM Component planes. In *Lecture Notes in Computer Science: Artificial Neural-Networks–ICANN 2007*; de Sá, J.M., Alexandre, L.A., Duch, W., Mandic, D., Eds.; Springer: Berlin, Germany, 2007; Volume 4669, pp. 379–388.

Birks, J.B.H. Overview of Numerical Methods in Paleolimnology. In *Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques*; Birks, J.B.H., Lotter, A.F., Juggins, S., Smol, J.P., Eds.; Springer: London, UK, 2012; Volume 5, pp. 35–86.

Brosse, S.; Giraudel, J.L.; Lek, S. Utilisation of non-supervised neural networks and principal component analysis to study fish assemblages. *Ecol. Model.* **2001**, *146*, 159–166.

Cheng, L.; Lek, S.; Lek-Ang, S.; Li, Z. Predicting fish assemblages and diversity in shallow lakes in the Yangtze River basin. *Limnologica* **2012**, *42*, 127–136.

Choi, B.Y.; Yun, S.T.; Kim, K.H.; Kim, J.W.; Kim, H.M.; Koh, Y.K. Hydrochemical interpretation of South Korea groundwater monitoring data using Self-Organizing Maps. *J. Geochem. Explor.* **2014**, *137*, 73–84.

Cinar, O.; Merdun, H. Application of an unsupervised artificial neural network technique to multivariate surface water quality data. *Ecol. Res.* **2009**, *24*, 163–173.

Cvetkoska, A.; Jovanovska, E.; Francke, A.; Tofilovska, S.; Vogel, H.; Levkov, Z.; Donders, T.H.; Wagner, B.; Wagner-Cremer, F. Ecosystems regimes and responses in a coupled ancient lake system from MIS 5b to present: The diatom record of lakes Ohrid and Prespa. *Biogeosciences* **2016**, *13*, 3147–3162.

Ejarque-Gonzalez, E.; Butturini, A. Self-Organising Maps and Correlation Analysis as a Tool to Explore Patterns in Excitation-Emission Matrix Data Sets and to Discriminate Dissolved Organic Matter Fluorescence Components. *PLoS ONE* **2014**, *9*, e99618, doi:10.1371/journal.pone.0099618.
Ferrente, M.; Oliver Conti, G.; Fiore, M.; Rapisarda, V. Harmful Algal Blooms in the Mediterranean Sea: Effects on Human Health. *EMBJ* **2013**, *8*, 25–34.

Goncalves, M.L.; Netto, M.L.A.; Costa, J.A.F.; Zullo Junior, J. An unsupervised method of classifying remotely sensed images using Kohonen self-organizing maps and agglomerative hierarchical clustering methods. *Int. J. Remote Sens.* **2008**, *29*, 3171–3207.

Gore, P.A. Cluster Analysis. In *Handbook of Multivariate Statistics and Mathematical Modeling*; Tinsley, H.E.A., Brown, S.D., Eds.; Academic Press: San Diego, CA, USA, 2000; pp. 297–322.

Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G.; Papastergiadou, E. Evaluating the contributing environmental parameters associated with eutrophication in a shallow lake by applying artificial neural networks techniques. *Fresen. Environ. Bull.* **2017**, *26*, 3200–3208.

Hadjisolomou, E.; Stefanidis, K.; Papatheodorou, G.; Papastergiadou, E. Assessing the Contribution of the Environmental Parameters to Eutrophication with the Use of the “PaD” and “PaD2” Methods in a Hypereutrophic Lake. *Int. J. Environ. Res. Public Health* **2016**, *13*, 764, doi:10.3390/ijerph13080764.

Hardle, W.; Simar, L. *Applied Multivariate Statistical Analysis*; Springer: Berlin/Heidelberg, Germany, 2003.

Hoeger, S.J.; Shaw, G.; Hitzfeld, B.C.; Dietrich, D.R. Occurrence and elimination of cyanobacterial toxins in two Australian drinking water treatment plants. *Toxicon* **2004**, *43*, 639–649.

Ignatiades, L.; Gotsis-Skretas, O. A review on Toxic and Harmful Algae in Greek Coastal Waters (E. Mediterranean Sea). *Toxins* **2010**, *2*, 1019–1037, doi:10.3390/toxins2051019.

Jeppesen, E.; Sondergaard, M.; Lauridsen, T.L.; Kronvang, B.; Beklioglu, M.; Lammens, E.; Jensen, H.S.; Köhler, J.; Ventelä, A.M.; Tarvainen, M.; et al. Danish and other European experiences in managing shallow lakes. *Lakes Reserv. Manag.* **2007**, *23*, 439–451.

Jeppesen, E.; Sondergaard, M.; Liu, Z. Lake Restoration and Management in a Climate Change Perspective: An Introduction. *Water* **2017**, *9*, 122, doi:10.3390/w9020122.
Kagalou, I.; Leonardos, I. Typology, classification and management issues of Greek lakes: Implication of the Water Framework Directive (2000/60/EC). *Environ. Monit. Assess.* **2009**, *150*, 469–484.

Kagalou, I.; Papastergiadou, E.; Leonardos, I. Long term changes in the eutrophication process in a shallow Mediterranean lake ecosystem of W. Greece: Response after the reduction of external load. *J. Environ. Manag.* **2008**, *87*, 497–506.

Kohonen, T. *Self-Organizing Maps*, 3rd ed.; Springer-Verlag::Heidelberg, Berlin, Germany, 2001, doi:10.1007/978-3-642-56927-2.

Koussouris, T.S.; Diapoulis, A.C.; Balopoulos, E.T. Assessing the trophic status of Lake MikriPrespa, Greece. *Ann. Limnol.* **1989**, *25*, 17–24.

Lee, B.-H.; Scholz, M. A comparative study: Prediction of constructed treatment wetland performance with k-nearest neighbors and neural networks. *Water Air Soil Pollut.* **2006**, *174*, 279–301.

Lee, B.-H.; Scholz, M. Application of the self-organizing map (SOM) to assess the heavy metal removal performance in experimental constructed wetlands. *Water Res.* **2006**, *40*, 3367–3374.

Leng, M.; Wagner, B.; Boehm, A.; Panagiotopoulos, K.; Vane, C.H.; Snelling, A.; Haidon, C.; Woodley, E.; Vogel, H.; Zanchetta, G.; et al. Understanding past climatic and hydrological variability in the Mediterranean from lake Prespa sediment isotope and geochemical record over the Last Glacial cycle. *Quat. Sci. Rev.* **2013**, *66*, 123–166.

Löffler, H.; Schiller, E.; Kusel, E.; Kraill, H. Lake Prespa, a European natural monument, endangered by irrigation and eutrophication? *Hydrobiologia* **1998**, *384*, 69–74.

Maier, H.R.; Dandy, G.C. Neural Network Based Modelling of Environmental Variables: A Systematic Approach. *Math. Comput. Model.* **2001**, *33*, 669–682.

Moustris, K.; Larissi, I.; Nastos, P.; Paliatsos, A. Precipitation Forecast Using Artificial Neural Networks in Specific Regions of Greece. *Water Resour. Manag.* **2011**, *25*, 1979–1993.

Mueller, U.A.; Grunsky, E.C. Multivariate spatial analysis of lake sediment geochemical data; Melville Peninsula, Nuvanut, Canada. *Appl. Geochem.* **2016**, *75*, 247–262.
Mutil, N.; Chau, K.-W. Machine-learning paradigms for selecting ecologically significant input variables. *Eng. Appl. Artif. Intell.* 2007, 20, 735–744.

Nastos, P.T.; Moustris, K.P.; Larissi, I.K.; Paliatsos, A.G. Rain intensity forecast using Artificial Neural Networks in Athens, Greece. *Atmos. Res.* 2013, 119, 153–160.

Odabas, M.S.; Leelaruban, N.; Simsek, H.; Padmanabhan, G. Quantifying impact of droughts on barley yield in North Dakota, USA using multiple linear regression and artificial neural network. *Neural Netw. World* 2014, 4, 343–355.

Oh, H.-M.; Ahn, C.-Y.; Lee, J.-W.; Chon, T.-S.; Choi, K.H.; Park, Y.-S. Community patterning and identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using artificial neural networks. *Ecol. Model.* 2007, 203, 109–118.

Panagiotopoulos, K.; Aufgebauer, A.; Schäbitz, F.; Wagner, B. Vegetation and climate history of the Lake Prespa region since the Last Glacial. *Quat. Int.* 2013, 293, 157–169.

Park, Y.-S.; Lek, S.; Scardi, M.; Verdonschot, P.; Jorgensen, S.E. Patterning exergy of benthic macroinvertebrate communities using self-organizing maps. *Ecol. Model.* 2006, 195, 105–113.

Park, Y.-S.; Verdonschot, P.; Chon, T.-S.; Lek, S. Patterning and predicting aquatic macroinvertebrate diversities using artificial neural network. *Water Res.* 2003, 37, 1749–1758.

Peeters, L.; Bacao, F.; Lobo, V.; Dassargues, A. Exploratory data analysis and clustering of multivariate spatial hydrogeological data by means of GEO3DSOM, a variant of Kohonen’s Self-Organizing Map. *Hydrol. Earth Syst. Sci.* 2007, 11, 1309–1321.

Pyrovetsi, M. Integrated Mediterranean Programmes and the Natural Environment: A case study in Greece. *Environmentalist* 1989, 9, 201–211.

Recknagel, F.; Talib, A.; van der Molen, D. Phytoplankton community dynamics of two adjacent Dutch lakes in response to seasons and eutrophication control unravelled by non-supervised artificial neural networks. *Ecol. Inform.* 2006, 1, 277–285.

Reimann, C.; Filzmoser, P.; Garrett, R.; Dutter, R. *Statistical Data Analysis Explained: Applied Environmental Statistics with R*; Wiley: West Sussex, UK, 2008.
Rub, G.; Kruse, R.; Schneider, M.; Wagner, M. Visualization of Agriculture Data Using Self-Organizing Maps. In Applications and Innovations in Intelligent Systems; Allen, T., Ellis, R., Petridis, M., Eds.; Springer: London, UK, 2009, Volume XVI, pp. 47–60.

Scavia, D.; Allan, D.; Arend, K.K.; Bartell, S.; Beletsky, D.; Bosch, N.S.; Brandt, S.B.; Briland, R.D.; Daloğlu, I.; DePinto, J.V.; et al. Assessing and addressing the re-eutrophication of Lake Erie: Central basin hypoxia. J. Great Lakes Res. 2014, 40, 226–246.

Simsek, H. Mathematical modeling of wastewater-derived biodegradable dissolved organic nitrogen. Environ. Technol. 2016, 37, 2879–2889.

Sneath, P.H.A.; Sokal, R.R. Numerical Taxonomy; Freeman: San Fransisco, CA, USA, 1973.

Stefanidis, K.; Kostara, A.; Papastergiadou, E. Implications of Human Activities, Land Use Changes and Climate Variability in Mediterranean Lakes of Greece. Water 2016, 8, 483, doi:10.3390/w8110483.

Stefanidis, K.; Papastergiadou, E. Influence of hydrophyte abundance on the spatial distribution of zooplankton in selected lakes in Greece. Hydrobiologia 2010, 656, 55–65.

Stefanidis, K.; Papastergiadou, E. Relationships between lake morphometry, water quality, and aquatic macrophytes, in Greek Lakes. Fresen. Environ. Bull. 2012, 21, 3018–3026.

Tota-Maharaj, K.; Scholz, M. Modeling of Permeable Pavements for Treatment of Urban Runoff using Self-Organizing Maps. Environ. Eng. Manag. J. 2013, 12, 2273–2287.

Tryfon, E.; Moustaka-Gouni, M. Species composition and seasonal cycles of phytoplankton with special reference to the nanoplankton of Lake MikriPrespa. Hydrobiologia 1997, 351, 61–75.

Tsai, W.-P.; Huang, S.-P.; Cheng, S.-T.; Shao, K.-T.; Chang, F.-J. A data-mining framework for exploring the multi-relation between fish species and water quality through self-organizing map. Sci. Total Environ. 2017, 579, 474–483.

Vardaka, E.; Moustaka-Gouni, M.; Cook, C.; Lanaras, T. Cyanobacterial Blooms and water quality in Greek waterbodies. J. Appl. Phycol. 2005, 17, 391–401.
Vesanto, J.; Alhoniemi, E. Clustering of the Self-Organizing Map. *IEEE Trans. Neural Netw.* **2000**, *11*, 586–600.

Vesanto, J.; Alhoniemi, E.; Himberg, J.; Parhankangas, J. *SOM Toolbox for Matlab*. 2000. Available online: http://www.cis.hut.fi/projects/somtoolbox/ (accessed on 2 May 2016).

Wang, B.; Li, H.; Sun, D. Social-Ecological Patterns of Soil Heavy Metals Based on a Self-Organizing Map (SOM): A Case Study in Beijing, China. *Int. J. Environ. Res. Public Health* **2014**, *11*, 3618–3638, doi:10.3390/ijerph110403618.

Yang, X.; Wu, X.; Hao, H.; He, Z. Mechanisms and assessment of water eutrophication. *J. Zhejiang Univ. Sci.* **B2008**, *9*, 197–209.

Zacharias, I.; Bertachas, I.; Skoulikidis, N.; Koussouris, T. Greek Lakes: Limnological overview. *Lakes Reserv. Res. Manag.* **2002**, *7*, 55–62.

Zhang, L.; Scholz, M.; Mustafa, A.; Harrington, R. Assessment of the nutrient removal performance in integrated constructed wetlands with the self-organizing map. *Water Res.* **2008**, *42*, 3519–3527.