# Machine Learning modeling techniques for forecasting the trophic state in a restored South Mediterranean lagoon using Chlorophyll-a in connection with the physico-chemical variables

Nadia Ben Hadid

# Doctoral THESIS in Cotutelle

To obtain the degree of

Doctor

**Delivered by**

**NATIONAL AGRONOMIC INSTITUTE OF TUNISIA &**
**UNIVERSITY OF PERPIGNAN VIA DOMITIA**

**Prepared within the doctoral schools: STAE & ED 305**
**And research units:**
Ecosystems and Aquatic Resources (UR03AGRO1)
Institute for Modeling and Analysis in Geo-environment and Health (IMAGES-ESPACE DEV)

Specialty : **Oceanography**

Presented by Ms. **Nadia BEN HADID**

**Machine Learning modeling techniques for forecasting the trophic state in a restored South Mediterranean lagoon (North Lagoon of Tunis) using Chlorophyll-*a* in connection with the physico-chemical variables.**

Publicly defended on **08 December, 2021** before the jury composed of

| | |
|---|---|
| **Mrs., Maria-Angela BASSETTI**, Professor, University of Perpignan Via Domitia, France. | **Chair & Examiner** |
| **Mr., Ezzedine MAHMOUDI,** Professor, Carthage, University, Bizerte, Tunisia. | **Reviewer** |
| **Mrs., Asma SAKKA HLAILI,** Professor, Carthage, University, Bizerte, Tunisia. | **Reviewer** |
| **Mrs., Evangelia KRASAKOPOULOU,** Professor, University of the Aegean, Greece. | **Examiner** |
| **Mrs. Catherine GOYET**, Professor, University of Perpignan Via Domitia, France. | **Thesis co-supervisor** |
| **Mr. Abdessalem SHILI,** Assistant Professor, National Agronomic Institute of Tunisia, Tunisia. | **Thesis co-supervisor** |

***I want to dedicate this work to my parents who stood by my side morally and financially through this long and difficult journey.***

***I could not have done it without you!***

# *Acknowledgement*

Foremost, I would like to express my sincere gratitude to my advisors Assistant Professor Abdessalem SHILI and Pr. Catherine GOYET for the continuous support of my PhD study and research, for their patience, motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and writing of this thesis.

Besides my advisors, I would like to thank the rest of my thesis committee: Pr. Maria-Angela BASSETTI, Pr. Ezzedine MAHMOUDI, Pr. Asma SAKKA HLAILI, and Pr. Evangelia KRASAKOPOULOU for accepting to assess my work and taking part in the accomplishment of this project.

Special Thanks to Assistant Professor Hatem CHAAR for his huge help. I am very grateful for the time that he took to work with me.

My thanks also go to Dr. Naceur BEN MAIZ, for offering me to work in Al-Buhaira Invest company to realize my thesis project.

My brother Mahmoud BEN HADID, I just love you, for everything you do, to make my life a bit easier.

Last but not least, I would like to thank my parents Saida BEN HADID and Jamal BEN HADID, for supporting me in every single step in my life and letting me feel special.

# *Abbreviations list*

ANN: Artificial Neural Network

ACF: Autocorrelation function

ADF: Augmented Dickey Fuller test

APHA: American Public Health Association

ARIMA: Autoregressive Integrated Moving Average

Chl-*a*: Chlorophyll-*a*

DO: Dissolved oxygen

EU: European Union

ML: Machine Learning

MSE: Mean squared Error

MSFD: Marine Strategy Framework Directive

MVLR: Multivariate Linear Regression

NARX: Nonlinear Autoregressive with external inputs neural network

OECD: Organization for Economic Development and Cooperation

PACF: Partial autocorrelation function

R: Coefficient of correlation

$R^2$: Coefficient of determination

RF: Random Forest

SARIMA: Seasonal Autoregressive Integrated Moving Average

SD: Secchi depth

TN: Total Nitrogen

TP: Total Phosphorus

WFD: Water Framework Directive

WHO: World Health Organization

# Abstract

Eutrophication episodes are commonly observed in coastal areas, causing significant damage to the ecosystem, especially in the Mediterranean Basin which represents the major world's tourism hotspot. For this reason, anticipate their presence, is a matter of importance, to prevent risks. Predictive models are effective techniques for eutrophication forecasting since ecologists and environmentalists are able to predict water pollution levels and take necessary precaution measures in advance. Previous studies have confirmed the superiority of Machine Learning (ML) in modeling water quality parameters.

In this work, a characterization, in terms of space, time and season of the physico-chemical water quality parameters in the North Lagoon of Tunis, a shallow restored Mediterranean coastal ecosystem located in the north of Tunisia is done. The present study of seven physico-chemical parameters (Secchi depth, dissolved oxygen, pH, salinity, water temperature, total phosphorus and total nitrogen), in addition to the Chl-*a* has covered approximately three decades (January 1989 - April 2018) without interruption, at five representative positions throughout the area of the North lagoon of Tunis. In this study, chlorophyll-*a* (Chl-*a*) is used as water quality indicator. After preprocessing the data, an Artificial Neural Network (ANN), a data-driven modeling approach from ML techniques, is proposed to predict the ecological state of the North Lagoon of Tunis. A Nonlinear Auto Regressive with eXternal inputs (NARX) neural network model was fitted to predict Chl-*a* concentrations in the North Lagoon of Tunis as an eutrophication indicator. The predictor variables (inputs) used are those that contribute the most to the variation of Chl-*a* concentration according to the Random Forest (RF) model, which is also a technique belonging to ML. According to RF, Secchi depth along with dissolved oxygen are the variables that most condition the variation of Chl-*a*. Various model scenarios with different NARX architectures were tested for Chl-*a* predictions. To verify the model performances, the trained models were applied to field monitoring data. In

addition, various mathematical parameters to measure the performance of the model were calculated (R, $R^2$, MSE). Results indicated that the developed NARX model can predict one step ahead (1 month) the Chl-*a* concentrations in the North Lagoon of Tunis with high accuracy (R= 0.79; $R^2$= 0.62; MSE= 0.31). In addition, results showed that RF and NARX models generally performed better than the multivariate linear regression ($R^2$= 0.2).

Besides the NARX network, a Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model was developed to forecast monthly Chl-*a* concentrations in the North Lagoon of Tunis, using three decades of Chl-*a* historical data. Results showed SARIMA $(2,0,2)(2,0,2)_{12}$ was the best fitted model for Chl-*a* forecasting in the North Lagoon of Tunis. The developed SARIMA model was validated with actual monthly Chl-*a* concentrations from last observations. Furthermore, we have demonstrated that with only one input variable SARIMA model shows great applicability as an eutrophication early warning tool using past Chl-*a* data. Finally, the developed SARIMA model was used to forecast Chl-*a* concentrations on a long-term scale for the period starting from May 2018 through December 2025, as a predictive approach to the ecosystem management that could serve for decision makers for the future generations. Based on the strong forecasting ability of the SARIMA model, and the powerful expression ability of the NARX neural network in nonlinear relationships, a hybrid model which combines SARIMA and NARX is also proposed in this study, for Chl-*a* forecasting. The use of linear and nonlinear approaches to model the linear and nonlinear terms (respectively) of Chl-*a* time series in the hybrid model increased the efficiency and accuracy of the predictions. The tests carried out for the hybrid approach proved its excellent performance. The use of hybrid model improved the prediction capability of Chl-*a* time series with the best performance being achieved (R= 0.82; $R^2$=0.68; MSE= 0.24).

For comparison, we simulated the same approaches on the historical data of the Station 5 (shown to be the most affected area by the eutrophication).

**Keywords:** Tunis North Lagoon, Eutrophication, Chlorophyll-*a*, Machine Learning, Predictive modeling, Forecasting.

## Résumé

Les épisodes d'eutrophisation sont couramment observés dans les zones côtières, causant des dommages importants aux écosystèmes littoraux, en particulier dans le bassin méditerranéen qui représente la principale destination du tourisme mondiale. Il est donc important d'anticiper leur manifestation pour prévenir les risques. Les modèles prédictifs sont des techniques relativement efficaces pour la prévision de l'eutrophisation, car les écologistes et les environnementalistes sont désormais capables de prévoir les niveaux de pollution de l'eau et de prendre les mesures de précaution nécessaires à l'avance. Des études antérieures ont confirmé la supériorité des algorithmes de l'apprentissage automatique (Machine Learning ou ML) dans la modélisation des paramètres de la qualité de l'eau.

Dans la présente étude, une caractérisation, en termes d'espace, de temps et de saison, de la qualité physico-chimique de l'eau dans la Lagune Nord de Tunis est établie. Cette lagune ayant fait l'objet d'un projet de restauration écologique, est un écosystème côtier méditerranéen, peu profond situé au nord de la Tunisie. Le suivi portant sur plusieurs paramètres physico-chimique (profondeur de Secchi, oxygène dissous, pH, salinité, température de l'eau, phosphore total et azote total), en plus de la Chlorophylle-*a* (Chl-*a*), a couvert environ trois décennies (Janvier 1989 - Avril 2018) sans interruption, dans ~~à~~ cinq stations représentatives de la Lagune Nord de Tunis. Dans cette étude, la Chl-*a* est utilisée comme indicateur de la qualité de l'eau. Après avoir réaliser le prétraitement des données, un réseau neuronale artificiel (ANN) qui est une approche issue des techniques ML est proposé. Un modèle de réseau neuronal non-linéaire autorégressif avec des entrées externes (Nonlinear autoregressive with external inputs neural network ou NARX) a été développé pour prédire les concentrations de la Chl-*a* dans la lagune en tant qu'indicateur d'eutrophisation. Les variables prédictives (entrées externes du NARX) utilisées sont celles qui contribuent le plus à la variation de la concentration de la Chl-*a* selon le modèle des forêts aléatoires (Random

Forest ou RF) qui est également une technique appartenant à ML. D'après RF, la profondeur de Secchi avec l'oxygène dissous sont les variables qui conditionnent le plus la variation de la Chl-*a*. Divers scénarios de modèles NARX avec différentes architectures ont été testés pour la prédiction de la Chl-*a*. Les simulations des divers modèles formés ont été comparées aux données réelles pour vérifier leurs performances. En plus, différents paramètres mathématiques ont été calculés (R, $R^2$, MSE). Les résultats ont indiqué que le modèle NARX développé peut prédire avec une grande précision (R=0,79 ; $R^2$=0,62 ; MSE= 0,31) la concentration de la Chl-*a* dans la Lagune Nord de Tunis avec un pas d'avance d'un mois. En outre, les résultats ont clairement confirmé que les modèles NARX et RF sont plus performants que la régression linéaire multivariée (MVLR) avec $R^2$=0,2. En plus du NARX, le modèle saisonnier, autorégressif, à moyenne mobile intégrée (SARIMA) a été développé pour prévoir les concentrations mensuelles de la Chl-*a*. Les résultats ont montré que SARIMA $(2,0,2)(2,0,2)_{12}$ est le modèle le plus adapté pour la prévision de la concentration de la Chl-*a* dans la Lagune Nord de Tunis. Le modèle SARIMA développé a été validé par les concentrations mensuelles réelles de la Chl-*a* relatives aux dernières observations. Il a été démontré qu'avec une seule variable d'entrée, le modèle SARIMA présente une grande applicabilité en tant que technique pour alerter en avance une manifestation d'eutrophisation dans l'écosystème. Enfin, le modèle SARIMA développé a été utilisé pour prévoir les concentrations de la Chl-*a* pour la période allant de Mai 2018 à Décembre 2025. Basé sur la forte capacité de prévision du modèle SARIMA, et de la puissante capacité d'expression du réseau neuronal NARX dans les relations non-linéaires entre les variables, un modèle hybride combinant SARIMA et NARX est proposé pour la prévision de la Chl-*a* dans cette étude. L'utilisation combinée (modèle hybride) des approches linéaires et non-linéaires pour modéliser les spécificités linéaires et non-linéaires (respectivement) de la série temporelle de la Chl-*a* a permis d'améliorer la qualité et la précisions des prévisions. L'utilisation du modèle

hybride a permis d'améliorer la capacité de prédiction d'un pas de temps en avance (1 mois) de la Chl-*a*, avec les meilleurs performances obtenues (R= 0,82; $R^2$= 0,67; MSE=0,24).

Pour comparaison, nous avons simulé les mêmes approches sur les données historiques de la station 5 (trouvée la plus affectée par l'eutrophisation).

**Mots-clés :** Lagune Nord de Tunis, Eutrophisation, Chlorophylle-*a*, Machine Learning, Modélisation prédictive, Prévisions.

# *Table of contents*

# *List of figures*

# *List of tables*

# General Introduction

Transitional water bodies, such as coastal lagoons, are situated at the interface between the continent and the sea. These are active zones that provide important ecological services (Mooney *et al*. 2009; Newton *et al*. 2018), and they cover around 13% of the world's coastline (Barnes, 1980). On a global scale, human activities have significantly affected coastal and transitional waters, making eutrophication of coastal ecosystems a worldwide issue, particularly in coastal lagoons (Nixon, 1995; Cloern, 2001).

Eutrophication is known as the enrichment of nutrients, composed of nitrogen and / or phosphorus, causing an accelerated development of phytoplankton and Macroalgae which causes an undesirable disturbance of the balance of organisms present in the water and a degradation of the quality of the water in question (Ferreira et al., 2011).

The increase in nutrient inputs, enhanced by urbanization, agriculture or industry, has led to complex direct and indirect responses by natural ecosystems (Schramm, 1999; Viaroli *et al*., 2008). Anoxic crises caused by a strong degradation of the organic matter, toxic algal blooms, loss of biodiversity, and more generally, deterioration of ecosystem functions and services are all consequences of anthropogenic eutrophication (Cloern, 2001; Zaldivar *et al*., 2008 a, b). Furthermore, this process can be a threat on the human health, for example, following the consumption of contaminated shellfish or fish (phycotoxins, ammonium).

Water Framework Directive (WFD) assumed that “Water is not a commercial product like any other but, rather, a heritage which must be protected, defended and treated as such.” (WFD, Directive 2000/60/EC).

Eutrophication is descriptor 5 of Marine Strategy Framework Directive (MSFD) and some criteria and methodological standards have been set for al ecosystems.

MSFD stated that those ecosystems constitute "a precious heritage that must be protected, preserved and, where practicable, restored with the ultimate aim of maintaining biodiversity and providing diverse and dynamic oceans and seas which are clean, healthy and productive" (MSFD, Directive 2008/56/EC).

For decades, Mediterranean coastal lagoons have been subjected to anthropogenic eutrophication, making them the most vulnerable ecosystems (Viaroli *et al*., 2005; Zaldívar *et al*., 2008; Souchu *et al*., 2010). They are impacted by highly inhabited and intensively farmed watersheds, particularly during the summer when the Mediterranean is a significant vacation destination (Vogiatzakis *et al*., 2006).

Aside from urban pressures, these habitats are vulnerable also due to their transitional state and their restricted exports to the open sea (de Jonge and Elliott, 2002; Newton *et al*., 2014).

Several studies have been undertaken to determine the extent of eutrophication in coastal Mediterranean lagoons. García-Ayllón (2017), stated that the Mar Menor lagoon, located in the east of the region of Murcia in Spain, has suffered an important process of intense anthropization over the last five decades. The rapid population growth of a new jellyfish species, which has reached over 100 million, especially in summer, was one of the principal indicators (Robledano *et al*., 2011). Thau Lagoon is another particularly interesting case of a Mediterranean coastal lagoon suffering from eutrophication. It is an ecosystem located at the Mediterranean French coast which is famous by supporting traditional shellfish farming activities in France. This lagoon has been subject to eutrophication leading to major anoxic events associated with massive mortalities of shellfish stocks (Derolez *et al*., 2020).

We can also mention the Palavasian lagoon complex, which is a collection of eight lagoons along the French Mediterranean coast that experienced extensive eutrophication over a four-decade period, mostly due to nutrient over-enrichment from constant sewage discharges (Leruste *et al*., 2016). The Ghar el Melh lagoon provides another good example for the study

of the eutrophication process in coastal Mediterranean lagoons. According to Shili *et al*. (2002), the lagoon experienced several dystrophic crises during the period 1994–1996. In addition, Turki *et al*. (2007) reported the proliferation of harmful algal species in the lagoon, such as *Kryptoperidinium foliaceum*, *Prorocentrum micans* and *Anabaena sp*.

The lagoon of Bizerte, on the south-western border of the Mediterranean Sea, located at the north of Tunisia is considered to have undergone eutrophication and has developed an intensive shellfish farming (Sakka Hlaili *et al*., 2008).

In this context, the North Lagoon of Tunis, a restored south Mediterranean lagoon in northern Tunisia, serves as a valuable example for eutrophication diagnosis and research in coastal ecosystems. In fact, this ecosystem has a long history of contamination and was one of the most polluted lagoons in the world (Harbridge *et al*., 1976; Afli *et al*., 2008; Armi *et al*., 2008). To limit the anthropogenic input, a major restoration project was implemented in 1985. Before this project was conducted, the above-mentioned lagoon was the main outlet for solid waste and domestic/industrial wastewaters stemming from the city of Tunis (the capital of Tunisia). Process such as dystrophic episodes, anoxia, fish mortality and red waters were observed in the North Lagoon of Tunis ecosystem (Belkhir, 1984; Zaouali and Batten, 1985).

The North lagoon of Tunis is currently a completely artificial environment following the human intervention, and the ecological follow-up is a necessity to guarantee the good ecological functioning of this ecosystem located in the center of the urban zone.

Chlorophyll-*a* (Chl-*a*) is the principal pigment in aerobic photosynthetic organisms. Its measurement is used to determine the amount of phytoplankton biomass in the water, and hence the extent of eutrophication in the environment (Tian *et al*., 2017). The probable presence of algae blooms that have a significant impact on the physical, chemical and biological processes of the lagoon can be interpreted as elevated Chl-*a* levels (Tian *et al*., 2017). According to the Organization for Economic Development and Cooperation (OECD,

1982), classification of trophic state, 8.00 μg $L^{-1}$ of Chl-*a* concentration is the threshold for eutrophication.

Cyanotoxins produced by cyanobacteria in lake water might endanger human health (Watzin *et al*., 2006; Mc Quaid *et al*., 2011; Kalaji *et al*., 2016). When the present concentration of cyanotoxins is unavailable, Chl-*a* is commonly used as a proxy metric for cyanobacterial density (Wheeler *et al*., 2012). Therefore, controlling Chl-*a* concentrations is essential and as a result providing information for water quality management.

The capacity to automatically monitor water quality is especially beneficial in sensitive locations where (1) there is a high risk of probable contamination episodes and (2) relevant socioeconomic activities are carried out that necessitate preventative actions. However, as far as anyone knows, there is no automated device that can correctly measure Chl-*a* concentrations in real time. Chl-*a* measurements must be done in laboratories or by satellites, which means high latency and high cost (Jimeno-Sáez *et al*., 2020), and when using satellites measurements, there is the common problem of missing data (Al Shehhi and Kaya, 2020). To avoid such inconveniences, most ecologists recently have been using modeling techniques.

The world is changing at an alarming rate, allowing study into ecosystems to become increasingly more complicated, and many new challenges are emerging every day (Griitzner, 1996). Some traditional methodologies, such as field investigations or laboratory experiments, are no longer sufficient for describing complex systems accurately and completely (Griitzner, 1996). Ecology has long acknowledged the necessity to collaborate with mathematics, statistics, informatics, and computational fields (Otto and Day, 2007). This interdisciplinary approach is well illustrated in some of the most frequently cited papers by early ecology pioneers such as Fisher, Preston and MacArthur, who used mathematical methods to describe and analyze ecological observations (Fisher *et al*., 1943; Mac Arthur, 1967).

The disciplines of mathematical and theoretical ecology have become a tool to ecologists to use. According to Codling and Dumbrell (2012), without these new disciplines in ecology, collecting data is a futile and meaningless task.

When an environmentalist's view is limited, a successful mathematical model relates the data to environmental problems and gives sufficient knowledge (Kompare *et al*., 1994). In terms of model development, there are typically two approaches in ecological modeling: (1) physically-based (or conceptual) and (2) data-driven-based models (Babovic, 2005; He *et al*., 2014; Zhang *et al*., 2016). On one hand, physically-based models identify the fundamental mechanisms for algae growth and outbreak (Hood *et al*., 2006; Zhang *et al*., 2013). However, owing to the complexities of marine environments and the vast range of parameters that need calibration, many physical, chemical and biological processes remain unclear. Such types of models are used mostly for scenario analysis rather than prediction (Recknagel *et al*., 2014). On the other hand, data-driven models are easier to incorporate, they are not so complex and eliminate the need for advanced knowledge of physical processes (Bowden *et al*., 2006). These models are popular and widely used for modeling complex natural processes, mostly in predictive modeling as predictive ability is considered by many to be the ultimate goal in ecology (Peters 1991).

Despite the number of studies that have been focusing on this eutrophication process in natural aquatic ecosystems (Delbaere and Nieto-Serradilla, 2004; García Pintado *et al*., 2007), there is still relatively few works done in setting up a proactive early-warning approach, to identify and prevent potential problems, especially in the Mediterranean scale.

Various statistical methodologies based on regression analysis have been used in predictive modeling. These traditional data processing methods, often employ a linear method to simplify complicated situations, resulting in poor outcomes since they are inefficient in dealing with complicated non-linear relationships between the variables involved (Su *et al*.,

2015). To overcome the mentioned restrictions, a productive and cost-efficient technique for eutrophication estimation using effective and resilient advanced approaches is required (Tiyasha *et al*., 2020). Machine learning (ML) methods have been demonstrated to be more efficient than traditional data processing techniques in forecasting water quality (Abba *et al*.,2017), owing to their ability to predict nonlinear and complicated functions. Previous research has shown that ML outperforms the conventional techniques in modeling water quality factors (Juntunen *et al*., 2012; Charulatha *et al*., 2017).

In this study we developed among ML techniques a commonly used neural network, in addition to the Random Forest model for feature selection, to forecast Chl-*a* concentrations in the North Lagoon of Tunis as an eutrophication indicator.

In predictive modeling and especially in forecasting tasks, when using time series, we need to work with the famous method of Box and Jenkins (1967) named Auto-regressive integrated moving average (ARIMA) model. This technique have simple structure and fast modeling features. It is also important to mention the SARIMA model, which consists of the ARIMA model, including the seasonal portion of time series data, which is very frequently used for monthly time series with seasonal patterns (Prista *et al*., 2011).

In time series modeling, a recent but commonly used practice for improving the forecasting accuracy has been made: the hybrid model. The hybrid model would take the best characteristics of the neural network and do the same for Box and Jenkins method, to perform a more accurate and reliable model. It is important to highlight that all of those techniques will be explained in further details in the following chapters.

The ecological stability of the North Lagoon of Tunis makes it of a significant socio-economic and ecological values. Multiples services are provided in this ecosystem, such as in tourism (water sports), in fisheries, and in the conservation of sea birds. Thus, it is necessary to improve our understanding of the eutrophication process and of the interactions among the

water quality parameters in the lagoon, to adopt sustainable management strategies. One of the most important indicator of the presence and degree of eutrophication in water bodies is the Chl-*a* content. Taking the North Lagoon of Tunis as the case study, the objectives of this thesis are divided in five chapters:

- In the first chapter, we will begin by presenting the study area. Then, we will characterize the spatio-temporal and seasonal variation of the physico-chemical parameters of the water in the North Lagoon of Tunis. To do so, monthly time series from January 1989 to April 2018 for each parameter (Chlorophyll-*a*, Secchi depth, total nitrogen, total phosphorus, pH, salinity, water temperature and dissolved oxygen) were used. The parameters were sampled at 5 representative Stations of the lagoon, set up according to a hydrodynamic model. This characterization is used to identify any hidden patterns in the water quality time series and choose the most adequate predictive modeling technique to evaluate and predict the eutrophication level in the North Lagoon of Tunis.
- In the second chapter, the objectives were: to use ML powerful nonlinear techniques (1) to select the specific variables that are most related to the Chl-*a* concentrations in the North lagoon of Tunis (2) to develop a predictive model to estimate and forecast one step ahead the Chl-*a* concentrations based on neural networks techniques, (3) to validate the performance of the predictive model. For comparison, the same approaches were performed on the historical data of Station 5 (the most affected area by the eutrophication, in the lagoon). Our findings in this chapter allowed us to write a paper published in *Wetlands Journal*.
- In the third chapter, the objective was to forecast on a long-term period the Chl-*a* concentrations as a eutrophication indicator, using approximately three decades of historical Chl-*a* data (January 1989 to April 2018) in the North Lagoon of Tunis,

simulating the technique of Box and Jenkins named Seasonal Auto-Regressive Integrated Moving Average (SARIMA) model. SARIMA algorithm was developed, and its accuracy discussed. For comparison, the same approach was performed on the historical data of Station 5. This chapter's results, enabled us to write a second paper submitted and under revision now in the *Journal of Coastal Conservation*.

- In the fourth chapter, a hybrid model was developed by combining the neural network with SARIMA technique to improve the forecasting accuracy of the Chl-*a* in the North Lagoon of Tunis. For comparison, the same approach was performed on the historical data of Station 5.
- The last chapter summarizes the main findings of this study and presents perspectives for future work.

# Chapter I: Study area presentation and physico-chemical parameters characterization

## 1. Introduction

Coastal lagoons are vital transition zones between freshwater and seawater, and as such, they are extraordinarily varied and complex structures (Levin *et al*., 2001; Pérez-Ruzafa *et al*., 2007). Nevertheless, they are subjected to natural and anthropogenic pressure. They are subject to intermittent or persistent seawater inputs during high tides and floods, and human impact on these ecosystems has significantly increased the quantity of entering nutrients (Nixon, 1995; Howarth and Marino, 2006).

The physico-chemical properties of these transitional zones are defined by the mixing of salt and fresh water, which varies substantially within and between annual cycles due to the unpredictability of the incoming fluxes. The emergence of substantial human populations in coastal locations over the previous centuries has changed the number and form of these fluxes, increasing their unpredictability and trophic charge (Elliott and Quintino, 2007).

The major issue impacting most of the world's coastal areas is eutrophication, which has serious and long-term consequences for ecosystems (Coelho *et al*., 2015). To assess the health of an ecosystem, it has been recognized that eutrophic symptoms (i.e., anoxia, high Chlorophyll-*a* concentrations, etc.) are primarily a result of eutrophication, and that measuring physico-chemical variables such as nutrients, salinity, dissolved oxygen, Chlorophyll-*a*, and so on can help understand these process (Coelho *et al*., 2015).

Population expansion, increased urbanization, and rural emigration to huge coastal agglomerations in Tunisia, as in numerous other countries in the Mediterranean's south and east, have resulted in a significant rise in water discharges, which is the cause of ecosystems degradation (Ben Maiz, 1997).

The current study was carried out at the North Lagoon of Tunis which provides a good example for this topic. It is one of Tunisia's most important lagoons, which has reached a critical ecological state as a result of urban growth. In fact, annual average salinity values exceeded 50 Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981) . The Chl-*a* annual average concentration was of about 61µg L-1 Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981). Nutrients showed high levels (the annual average concentration of the total phosphorus was equal to 600 µg L-1 and total nitrogen was equal to 4400 µg L-1 Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981). The dissolved oxygen rate varied between 0 to 200% Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981).

Its strategic location in the heart of the capital, along with its ecological significance, piqued the Tunisian government's interest, leading to the construction of a lagoon restoration project based on water circulation, which was held in 1985 and monitored by "Al Buhaira Invest company", which had the name of "Promotion company of Tunis Lagoon" or "Société de Promotion du Lac de Tunis" (SPLT) back then.

In this first chapter, we shall introduce the studied area. Then, we will look at the spatiotemporal and seasonal variations in the physico-chemical water quality parameters in the North Lagoon of Tunis.

To accomplish so, we employed a monthly time series from January 1989 to April 2018 (about three decades) of seven physico-chemical and one biological parameters (Secchi depth, total nitrogen, total phosphorus, pH, salinity, dissolved oxygen, temperature and Chlorophyll-*a*). These parameters were sampled at 5 representative Stations of the lagoon, set up according to the hydrodynamic model, established during the restoration project, mentioned earlier.

## 2. Study Area

The North Lagoon of Tunis is a well-mixed shallow coastal seawater lagoon located in the north of Tunisia (36°45′–36°52′ N and 10°10′–10°20′ E) and at the south of the Mediterranean Sea (Figure 1). Covering about 22 $km^2$, with an average depth of 2 m (range; 0.5 – 3.5 m), this lagoon is one of the largest shallow water bodies of the Tunisian coast (Trabelsi-Bahri, 2013). It is connected to the open sea (Gulf of Tunis) and water is exchanged with the Mediterranean Sea through the Kheireddine channel, which measures 800 m in length and 40 m in width and has a mean depth of approximately 2.5 m (Ben Charrada, 1992).

In 1985 a large restoration project had been undertaken in this lagoon to stop pollution and eutrophication (Van Berk and Oostinga, 1992). In this contaminated lagoon, the ultimate objective of this project was to achieve a good chemical and ecological status and to achieve significant land reclamation all around.

The restoration project resulted in a clear improvement of the biodiversity (Shili, 2008). However, being aware of the importance and fragility of this ecosystem, it must always remain under observation.

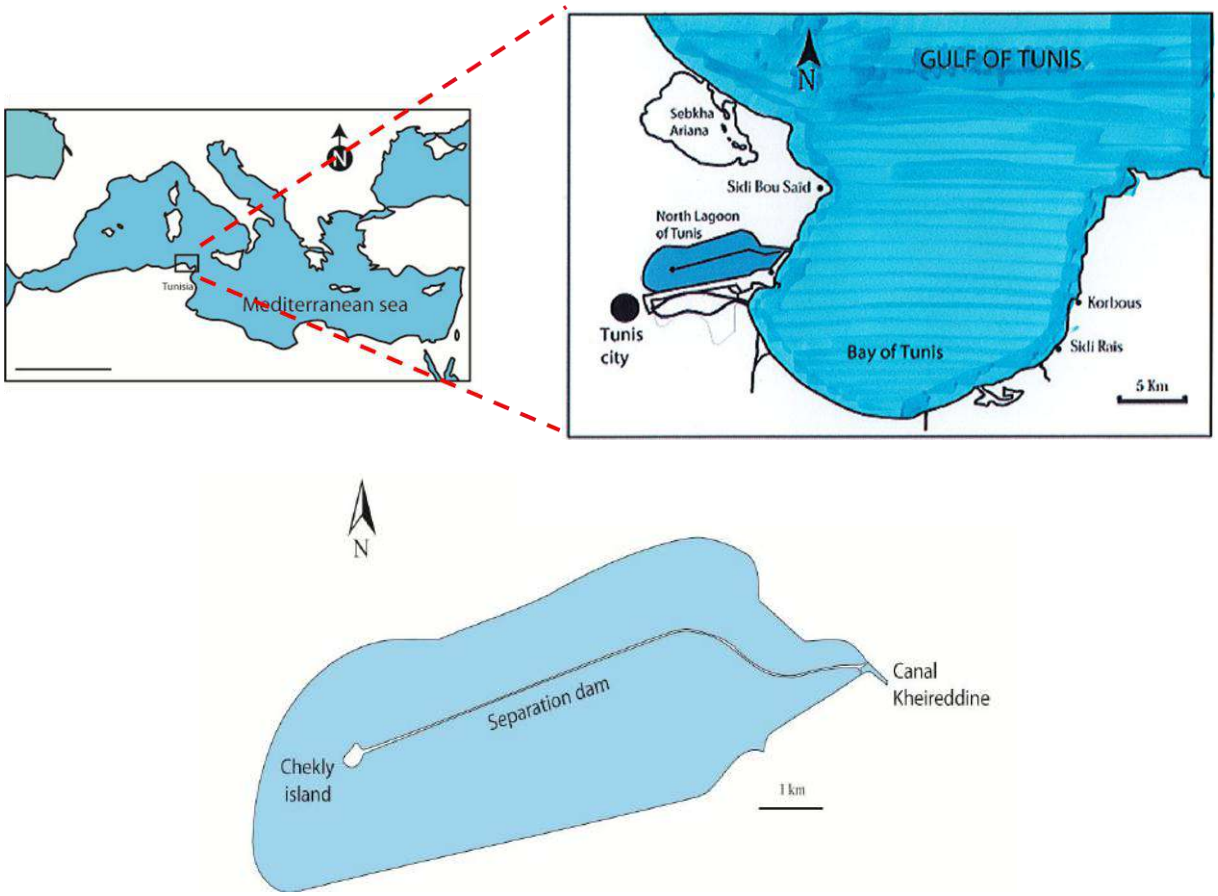



**Figure 1. Geographic location of the North Lagoon of Tunis.**

## 2.1. The history of the North Lagoon of Tunis genesis

The North Lagoon of Tunis has interested numerous researchers, historians, and geographers due to its crucial geographic location. According to Pimienta (1959), The North Lagoon was formed by the collapse of a Pliocene substratum. The geomorphologic development that resulted in its formation is presented in four major stages (Figure 2):

- A Holocene transgression allowed the transformation of a continental depression into a bay widely open to the sea (Zaouali, 1974).

- The development of a sedimentary coastline progressing towards the Southwest, between Rades and Kheireddine (Zaouali, 1974).

- Between the 5th and the 13th century BC, the coastline was developed in a fragmentary way between Rades and Kheireddine (Zaouali, 1974).

- According to Jouirou (1982), as a result from the activity of the Medjerda wadi, a deltaic transport from North to South, has led to a progressive separation between the lagoon and the sea by the formation of a littoral arrow (a double North-South tombolo); Thus, the North lagoon of Tunis, took shape in the 16th century.

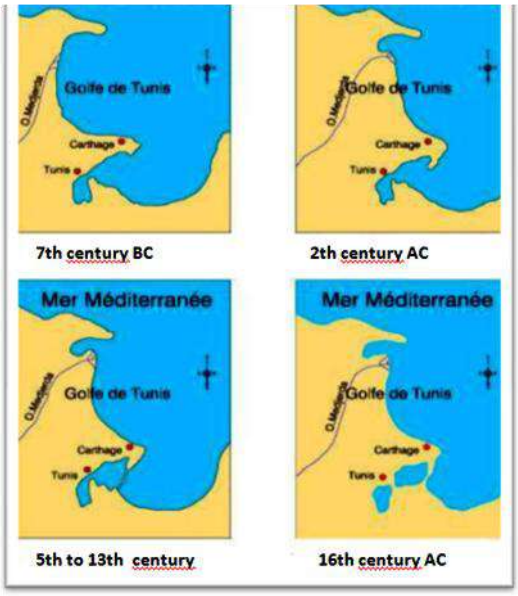



**Figure 2. The stages of the genesis of the Tunis lagoon (Pimienta, 1959).**

### 2.1.1. The degradation of the natural environment of the North Lagoon of Tunis

The first evidence of pollution in the North Lagoon of Tunis date back to 1724, when Peysonnel detected a siltation of the ecosystem as a result of continuous and increasing pollution (Zaouali, 1983). Indeed, since the lagoon's existence, effluent from the city has been rejected.

Eutrophication was observable by a very apparent ecological imbalance in the summer period (Figure 3), affecting the biological diversity: only a few tolerant fish species remained, such as *Mugil Ramada* (Zaouali, 1977) and a dozen of nitrophilic macro algae such as *Ulva*

(Zaouali et Beaten, 1984; Ben Maiz, 1995). This ecosystem deterioration led to the proliferation of sulfurous bacteria, anoxia and the appearance of red water (Zaouali, 1974; Schneider, 1977; Caumette, 1987; Van Berk et Oostinga, 1992).

The development of nitrophilic algae was favored by the shallow depth, abundant nutrients, adequate light, and weak water currents (Schneider, 1977).

In early summer, algae such as *Ulva* developed a considerable biomass in several areas of the lagoon disturbing, the water circulation (Belkhir et Hadj Ali Salem, 1981). By the end of the summer, rising temperatures cause widespread death of these algae. Decomposition necessitates a considerable consumption of dissolved oxygen, promoting the expansion of sulphate-reducing bacteria that produce sulfurous hydrogen (Zaouali; 1977; Caumette, 1987).

The presence of sulfurous hydrogen in addition to the lack of oxygen has made the environment toxic (Zaouali; 1977; Caumette, 1987).

**Figure 3. North lagoon of Tunis ecological situation before the restoration work**

**(On the right: Dead water zone, on the left: Biomass accumulation).**

## 2.2. The restoration project

In order to stop pollution and eutrophication, a massive restoration operation was launched in this lagoon in 1985 (Van Berk and Oostinga, 1992). The ultimate objective of this project was to attain good chemical and ecological status in this toxic environment, as well as to allow considerable land recovery all around. The goal was the reduction of water retention time in the lagoon (Figure 4). In order to do this, a tide-driven circulation system was developed after the construction of the (east-west) longitudinal dividing dam across the lagoon and the inlet/outlet gates at the entrance of the canal (Figure 5) connecting the lagoon to the open sea (Trabelsi-Bahri, 2013). The gates and the separation dam allowed strong circulation of the lagoon's water (Van Berk and Oostinga, 1992). In addition, the coastline was rectified to a straight line to prevent water stagnation (Trabelsi-Bahri, 2013).


**Figure 4. Water circulation in the North Lagoon of Tunis before restoration**

**Bleu arrows: lagoon waters renewal; Red arrows: arrival of sewage (Ben Maiz, 2008).**

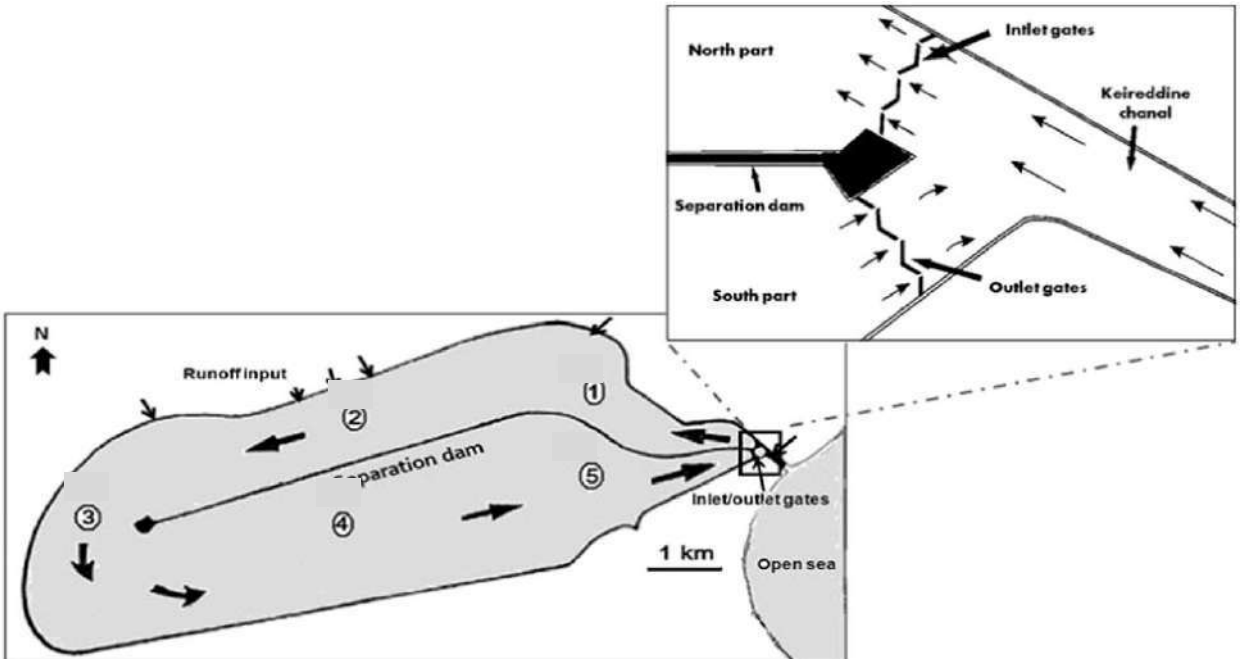



**Figure 5. The unidirectional inlet/outlet water circulation system after restoration of the North Lagoon of Tunis (Trabelsi-Bahri, 2013).**

## 2.3. The North lagoon of Tunis after the restoration works

After the completion of the restoration works, the ecosystem marked an improvement in water quality with the disappearance of *Ulva* in favor of filamentous algae *Chaetomorpha* and the expansion of marine magnoliophyta (Shili, 1995).

The average speed of the water streams became homogeneous with a reduced residence time ranging from 17 to 21 days, while before the restoration it was close to 28.5 to 30 days (Rezgui *et al*., 2008). After the restoration works, the lagoon was receiving about 1.6 million $m^3$ of seawater per day (Shili, 1995).

The quality of the water recorded in the lagoon after the restoration works showed a significant improvement (Table 1) with a marked reduction in eutrophication factors such as:

- The reduced fluctuations in salinity and pH of the water.
- The decrease in nutrient content in water (Nitrogen, Phosphorus).
- The absence of large fluctuations of dissolved oxygen.

- The absence of anoxia phenomenon in summer season.
- Significant improvement in the bacteriological quality of the waters: results of the analysis testify to water suitable for bathing in almost all areas of the lagoon.
- Significant improvement in water transparency: visibility often reaching down to the bottom.

**Table 1. Characteristics of the North lagoon of Tunis before and after the restoration work.**

| Variable | Before restauration work | After restauration work |
|---|---|---|
| **Salinity**<br><br>**Annual average** | **Value:** 28 to 50<br>**Reference:**<br>Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981) | **Value:** 32 to 43<br>**Reference:**<br>Ben Maiz (1992) ; Ben Charrada (1992);<br>Trabelsi – Annabi, 2001) |
| **Total nitrogen**<br><br>**Annual average ($\mu gL^{-1}$)** | **Concentration:**<br>4.400<br>**Reference:**<br>Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981) | **Concentration:**<br>460<br>**Reference:**<br>Ben Maiz (1992) ; Ben Charrada (1992);<br>Trabelsi – Annabi, 2001) |
| **Total phosphorus**<br><br>**Annual average ($\mu gL^{-1}$)** | **Concentration:**<br>600<br>**Reference:**<br>Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981) | **Concentration:**<br>20<br>**Reference:**<br>Ben Maiz (1992) ; Ben Charrada (1992);<br>Trabelsi – Annabi, 2001) |
| **Dissolved oxygen**<br><br>**Annual average (%)** | **Value:** 0 to 200<br>**Reference:**<br>Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981) | **Value:** 30 to 110<br>**Reference:**<br>Ben Maiz (1992) ; Ben Charrada (1992);<br>Trabelsi – Annabi, 2001) |
| **pH** | **Value:** 6.4 to 9.5<br>**Reference:**<br>Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981) | **Value:** 7.9 to 8.7<br>**Reference:**<br>Ben Maiz (1992) ; Ben Charrada (1992);<br>Trabelsi – Annabi, 2001) |
| **Chlorophyll-*a***<br><br>**Annual average ($\mu gL^{-1}$)** | **Concentration:**<br>61<br>**Reference:**<br>Zaouali (1974), Schneider (1978), Belkhir (1980), Belkhir et Hadj Ali (1981) | **Concentration:**<br>4<br>**Reference:**<br>Ben Maiz (1992) ; Ben Charrada (1992);<br>Trabelsi – Annabi, 2001) |

The restoration project resulted in a clear improvement of the lagoon (Ben Maiz, 1992). Thus, in January 2013, the North Lagoon of Tunis was deemed as a Wetland of International Importance, Ramsar site (Mdaini *et al*., 2019).

# 3. Methodology

The monthly concentrations of Chl-*a* data along with physico-chemical parameters of water quality of the North Lagoon of Tunis for the period from January 1989 to April 2018 were collected.

In the present study, a set of seven environmental variables known to affect Chl-*a* concentrations were monitored: Secchi depth, dissolved oxygen, total phosphorus, total nitrogen, pH, salinity and water temperature. Sampling cruises have been carried out from February 2014 through April 2018 at five sampling Stations. Water samples were collected about 10-20 cm below the water surface according to standard methods.

In addition, Al-Buhaira Invest Company, which is in charge of the ecosystem, provided us with important monthly time series sequences, as a part of the monitoring program for the lagoon, in order to gather information on the physical and chemical characteristics of the ecosystem.

## 3.1. Sampling cruise

The lagoon is characterized by its shallow depth; the sampling cruise were carried out using a coastal boat (Figure 6), with a flat bottom, of about 12 m in length and equipped with a 40-horsepower engine (Figure 7).

**Figure 6. The boat used in the field work.**


**Figure 7. The engine.**

## 3.2. Description of the Stations

For the coherence of the study, the sampling Stations are the same as those set up by the Al Buhaira Invest Company for the control of water quality. The Stations were chosen according to the hydrodynamic model established during the restoration project mentioned above (Figure 8).

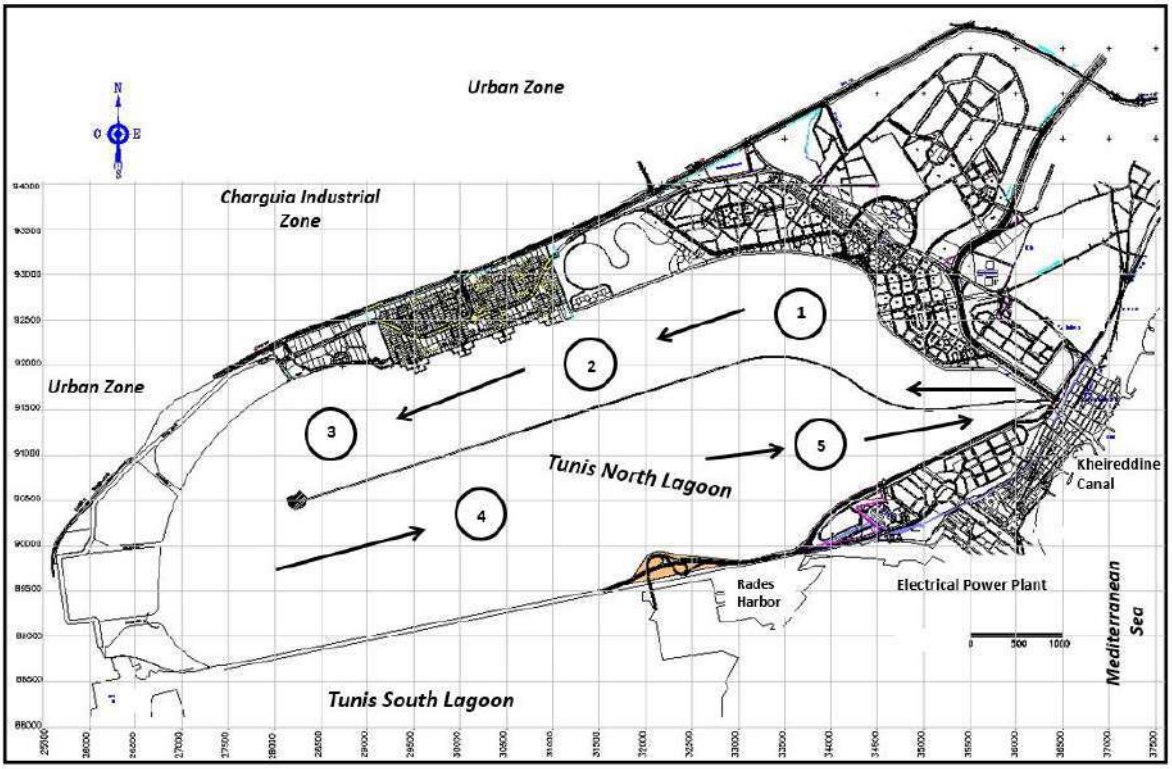



**Figure 8. Water quality monitoring Stations (1–5) in the North Lagoon of Tunis.**

The lagoon has been subdivided into five different compartments (Table 2).

**Table 2. Characteristics of the sampling Stations in the North lagoon of Tunis.**

| Stations | Characteristics |
|---|---|
| **Station 1** | It is the first section of the lagoon to receive seawater, which enters through the north gates of Kheireddine channel. This area's chemical, physical, and biological qualities are the most similar to those of the marine environment. |
| **Station 2** | It is a transitional environment between lagoon and marine ecosystems. |
| **Station 3** | It was the most damaged section prior to the restoration project. A significant amount of organic materials has been removed as part of the renovation procedure. |
| **Station 4** | It is the most important area representative of the South part. It is characterized by a shallow depth as well as a high concentration of macroalgae such as Chaetomorpha. |
| **Station 5** | It is the last area in the lagoon, the farthest from the marine influence. |

## 3.3. Physico-Chemical variables

Physico-chemical variables, including water temperature and practical salinity (Sp; IOC *et al*., 2010) called salinity, were measured *in situ* using a WTW LF325 conductivity meter (Figure 9). pH was measured by a pH 330i WTW pH meter.

The Secchi depth of the lagoon was measured with a 25 cm diameter Secchi disc (Figure 10). the Secchi depth is the visibility of the Secchi disc. In fact, the Secchi depth is a parameter indicator of the transparency of the water column and it is the depth of disappearance of the Secchi disc. In each Station we took two measurements of the Secchi depth.

Since the end of the restoration works, the visibility of the lake bottoms has improved significantly. In the absence of strong wind the rapport (transparency / depth) generally exceeds 90% (Shili, 1995).

Dissolved oxygen was measured by profiline OXY 197 oxymeter.

**Figure 9. Conductivity meter.**

**Figure 10. Secchi disc.**

Surface water samples were analyzed in the laboratory for total phosphorus and total nitrogen. Samples for nutrient determinations were collected in 1000 ml acid-washed polypropylene bottles and stored on ice (Figure 11).

**Figure 11. Preparation of samples for the analysis in the laboratory.**

Nutrient analyses were performed using spectrophotometric methods (Strickland and Parsons, 1972) with a UV–visible spectrophotometer. Total phosphorus TP and Total nitrogen TN were determined after alkaline peroxodisulfate digestion in an autoclave using unfiltered water. For the determination of total phosphorus, the phosphorus compounds were mineralized to orthophosphate ions in an autoclave at 100 °C using a solution of sulfuric acid and potassium persulfate.

Determination of total Nitrogen compounds requires high oxidation of the nitrogenous ions into nitrates, in an autoclave using an alkaline solution of persulfate, then by the reduction of nitrates to nitrites by passing through a cadmium column. The nitrites formed were determined using sulfanilamide and N-naphthyl-ethylene.

## 3.4. Chlorophyll-*a*

Water samples of 500 ml, for Chl-*a* measurement, were collected at 20-30 cm below the surface, and filtered through a whatman filter GF/C with a 0.45µm pore-size membrane (Millipore) and a diameter of 47 mm as shown in figure 12. We always try to proceed to the

analyses right away after the field work, but when it is impossible to do so, the samples are stored in the freezer at -20°C.

Nevertheless, After the filtration, the filters were stored at - 20°C until analysis. Chl-*a* was extracted in 10 ml of 90% acetone for 24 hours, in the dark at 4°C following the procedure given by Parsons *et al*. (1984). The extract concentration was analyzed spectrophotometrically (Figure 13) according to the method of Lorenzen (1967):

$$[\text{Chl-}a]\mu g\ L\text{-}1 = 26.7\ (A^{na}665 - A^{a}665)*v/(l*V) \quad (1)$$

With,

V= Filtred sea water (l)

v= Volume of extraction solvent (ml)

l= length of the cell used in the spectrophotometer

$A^{na}665$= Absorbance of the unacidified extract measured at 665 nm

$A^{a}665$= Absorbance of the acidified extract measured at 665 nm

**Figure 12. Water filtration device.**

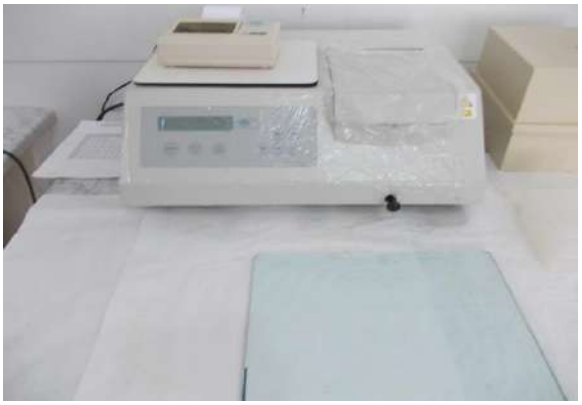


**Figure 13. The spectrophotometer.**

## 3.5. Analysis of variance (ANOVA)

Analysis of variance ANOVA was performed to ascertain if there was any significant difference in physico-chemical conditions (Total Nitrogen, Total Phosphorus, dissolved oxygen, Secchi depth, pH, water temperature, and salinity) and in Chl-*a* concentrations among the sampling Stations in the North Lagoon of Tunis.

ANOVA modeling was performed using the MATLAB software MATLAB® software (version 9.3.0.948333 (R2017b), The Mathworks, MA, USA).

# 4. Spatio-temporal and seasonal variation of the physico-chemical parameters in the North Lagoon of Tunis

Measuring physico-chemical characteristics is of a significant importance for assessing environmental quality in aquatic habitats. Often, physico-chemical characteristics condition the species distribution in the ecosystems.

It is also worth noting that in natural ecosystems, the eutrophication episodes may be enhanced by seasonal occurrences such as rainfall runoff, which modifies the physical and chemical properties of water due to the freshwater influx (Herrera- Silveira *et al*., 2002; Coelho *et al*., 2007). In addition, water temperature is one of the essential parameters that influences the growth of flora and fauna (Brown, 1992), hot and dry episodes can affect the natural ecosystem. Seasonal scales would be extremely helpful in analyzing the effects of climate change on ecological functions.

## 4.1. Results

### 4.1.1 Water temperature

Water temperature is an essential physical parameter in the circulation of water bodies as well as in biological cycles such as reproduction and dispersion of living organisms (Amniot et Chausepied, 1983).

It is a parameter that is affected by climatological variables. Because of the short depth of the North Lagoon of Tunis, water temperature varies as a function of the ambient air. The surface water temperature measured during daytime in the study sites ranged from 20 to 30 °C and 8

to 20.5 °C in spring/ summer and autumn/winter seasons, respectively. The temperature of the water is relatively homogeneous among the five Stations (Figure 14).

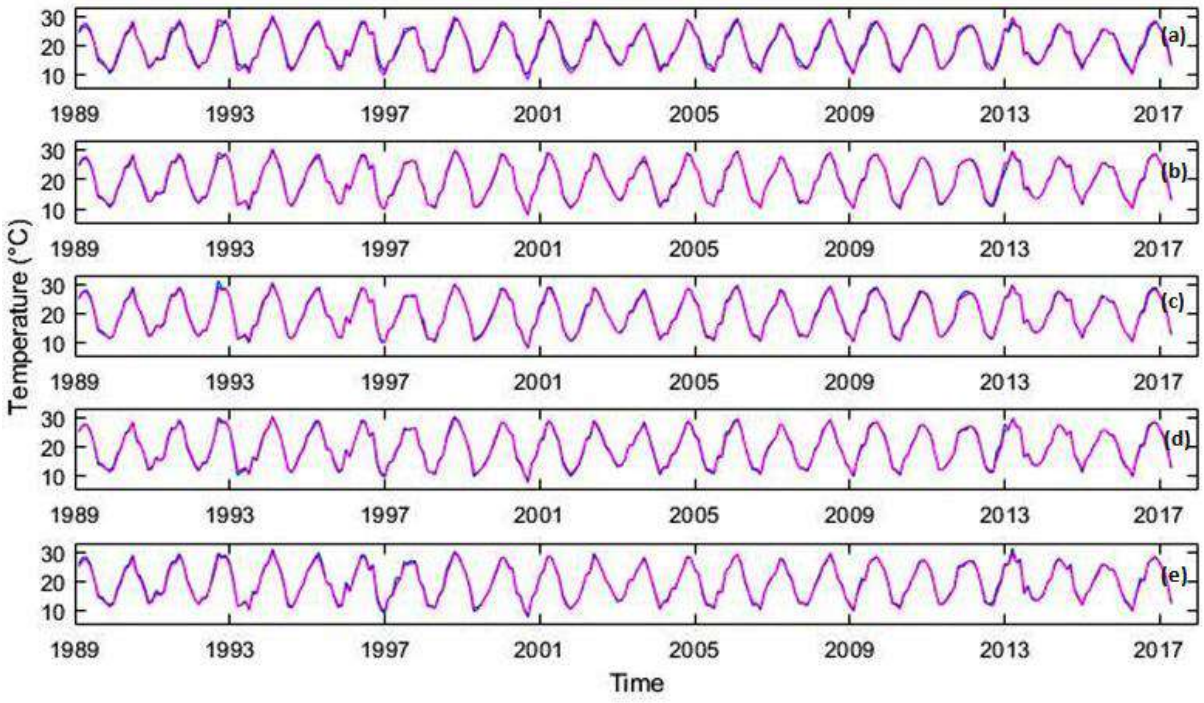



**Figure 14. Temporal variability of water temperature in the North Lagoon of Tunis: a) Station 1; b) Station 2; c) Station 3; d) Station 4; e) Station 5; blue line: temperature values at the Station - purple line: lagoon mean values over the sampling Stations.**

Clearly, the water temperature at the North Lagoon of Tunis fluctuates seasonally, due to the lagoon's shallowness (Figure 15).

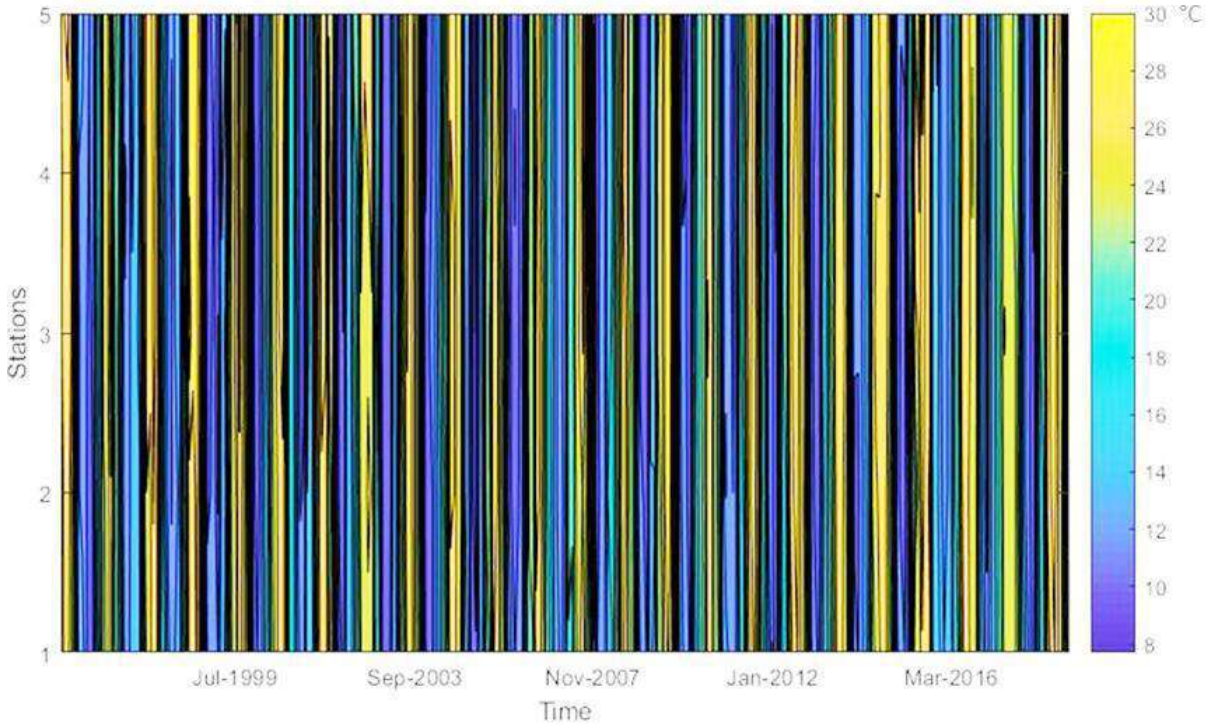



**Figure 15. Seasonal variation of the water temperature in the North Lagoon of Tunis.**

### 4.1.2 pH

During our study period, the mean pH values ranged from 7.2 to 8.7, in the North Lagoon of Tunis. The maximum pH noticed in the study period was 8.7 at Station 5 and the minimum pH recorded was 7.2 at Stations 1, 2 and 3 (Figure 16).

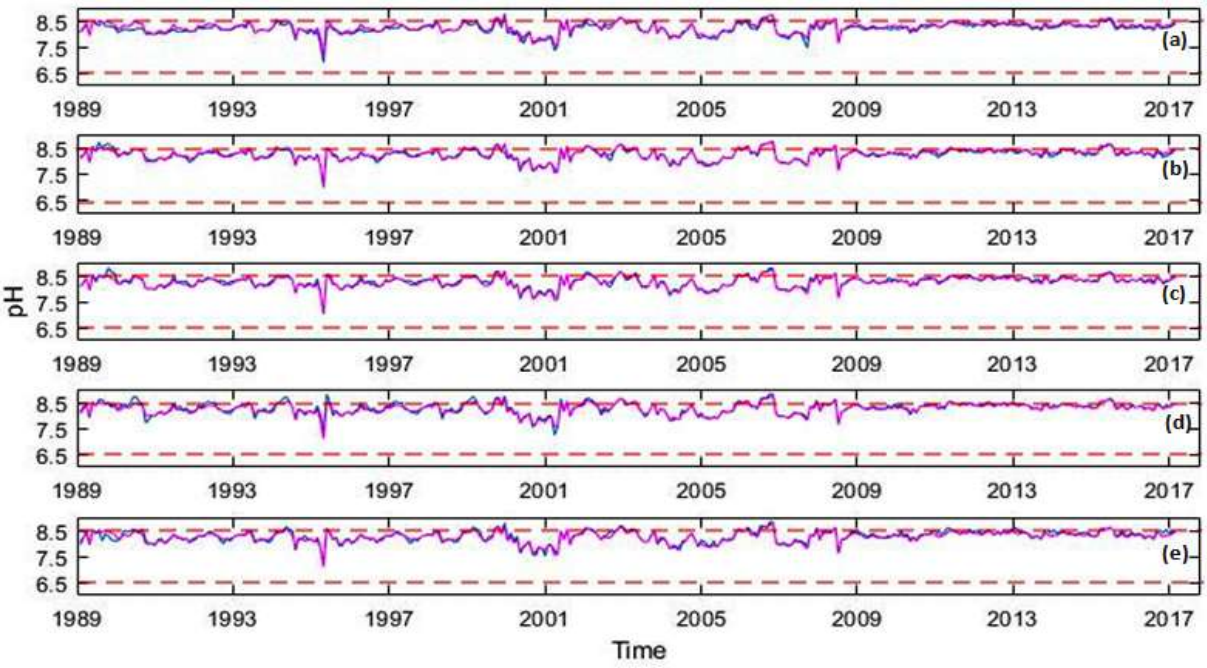



**Figure 16. Temporal variability of pH in the North Lagoon of Tunis: a) Station 1; b) Station2; c) Station 3; d) Station 4; e) Station 5; blue line: pH values at the Station - purple line: lagoon mean values over the sampling Stations - Red lines: APHA thresholds.**

Although the North Lagoon of Tunis was slightly alkaline throughout the year, it reached its lowest values in the winter and its highest values in the summer (Figure 17).

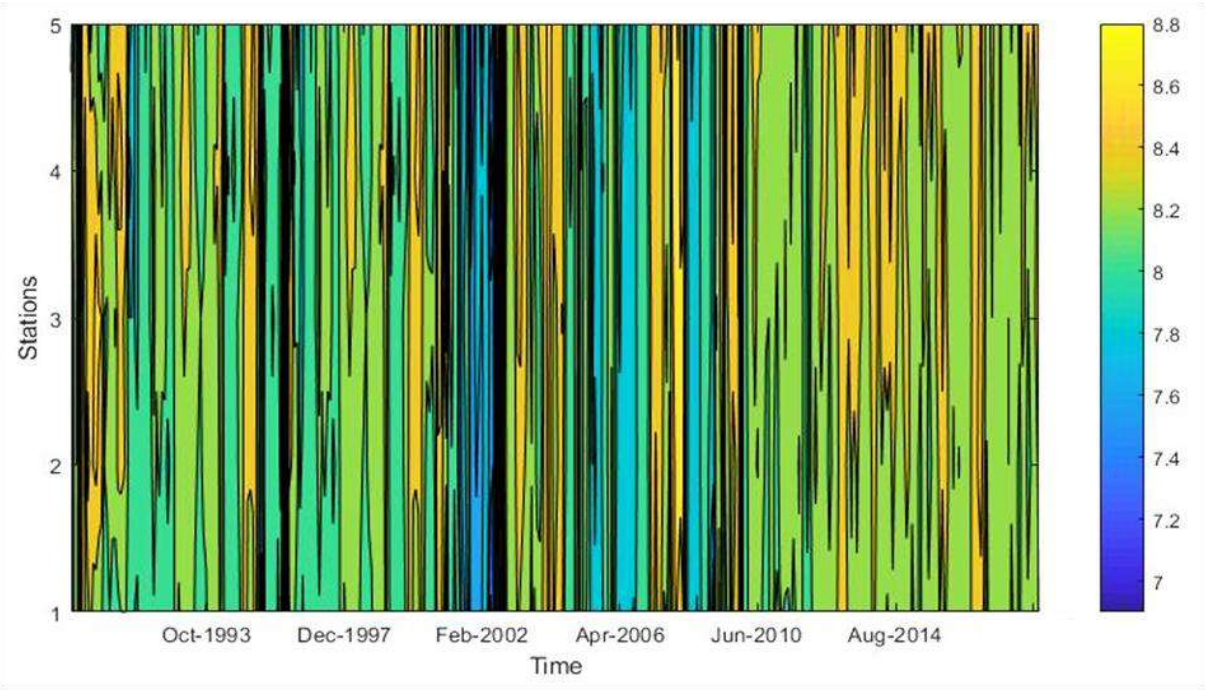



**Figure 17. Seasonal variation of the pH in the North Lagoon of Tunis.**

### 4.1.3 Secchi depth

Figure 18 shows the variation of Secchi Depth (SD) values obtained at the five Stations throughout the study period of observations. From June 1995 to June 2010, the deepest sites of the lagoon were Stations 3 and 5, while the shallowest was Station 4 throughout the research period. During the latter years of our observations, Station 3 became the deepest area in the lagoon due to some dredging operations.

The lowest depth measurements were at Station 4 (0.58 m) and the deepest measurements were at Station 3 (3.5 m).

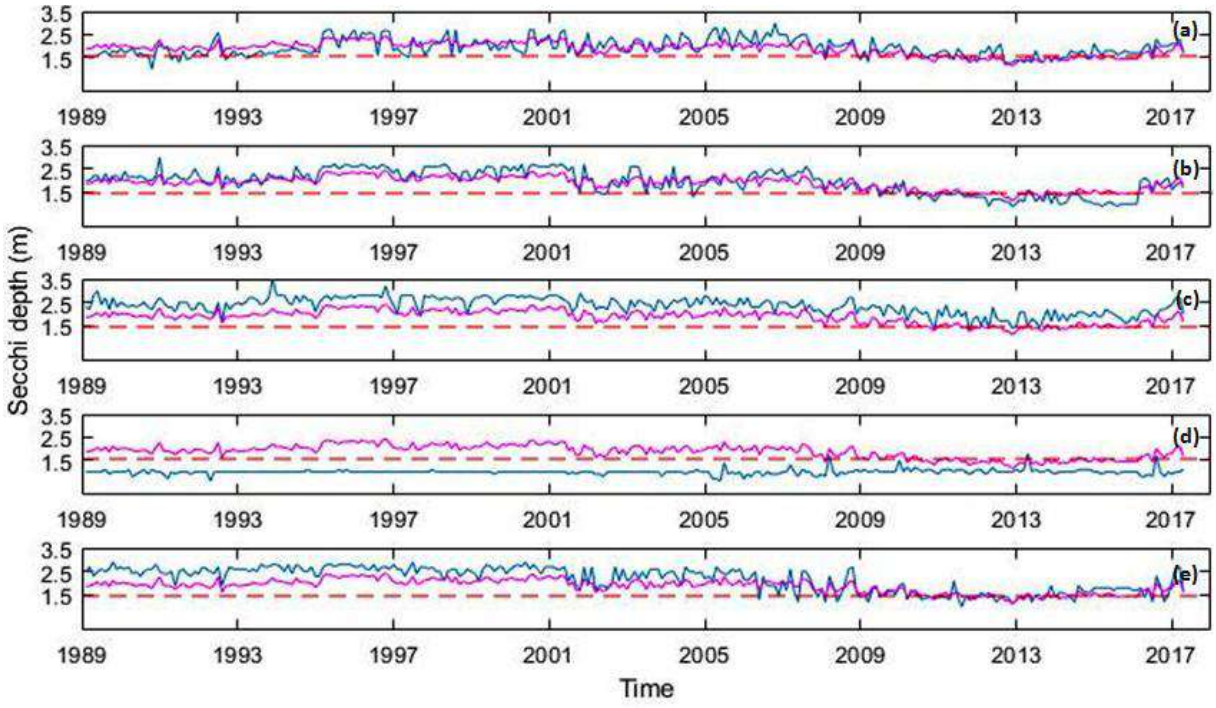



**Figure 18. Temporal variability of the Secchi depth in the North Lagoon of Tunis: a) Station 1; b) Station 2; c) Station 3; d) Station 4; e) Station 5; blue line: Secchi depth values at the Station - purple line: lagoon mean values over the sampling Stations - Red line: OECD threshold.**

The comparison of SD variations revealed that the seasonal fluctuations were quite apparent (Figure 19). In general, the SD was found to be shallowest during the summer and deepest during the winter at all sample sites. This can be attributed to water evaporation and a reduction in precipitation during the dry season (summer), and vice versa.

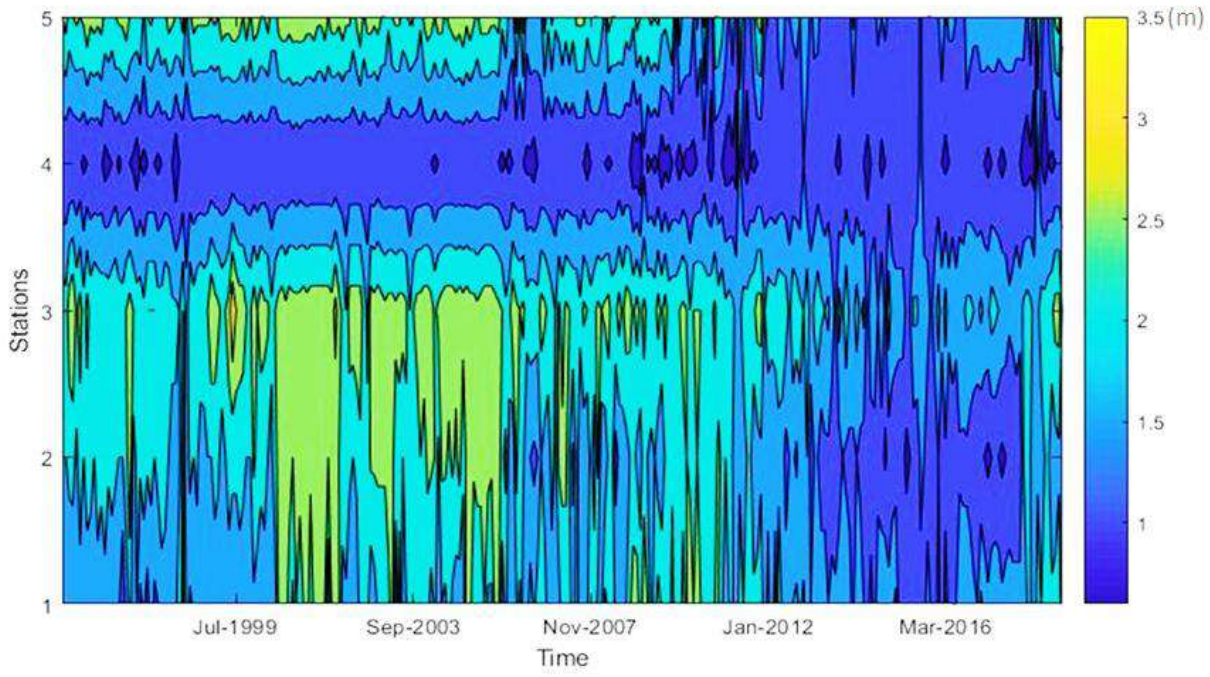



**Figure 19. Seasonal variation of the Secchi depth in the North Lagoon of Tunis.**

### 4.1.4 Total phosphorus

The total phosphorus (TP) is a fundamental factor in the fertility and productivity of the ecosystems (Sujitha and Ravindhranath, 2017).

Concentration of TP fluctuated significantly during the study periods. On one hand, the higher values were recorded at Station 5 (65 µg $L^{-1}$; above the lagoon average), especially in the wet season. These waters come from an area (Station 4) not affected by the dredging work undertaken as part of the lagoon restoration project, characterized by a very rich substrate in nutrients. On the other hand, the lowest phosphorous concentrations were measured at the second Station (4 µg $L^{-1}$) during the dry or hot season (Figure 20).

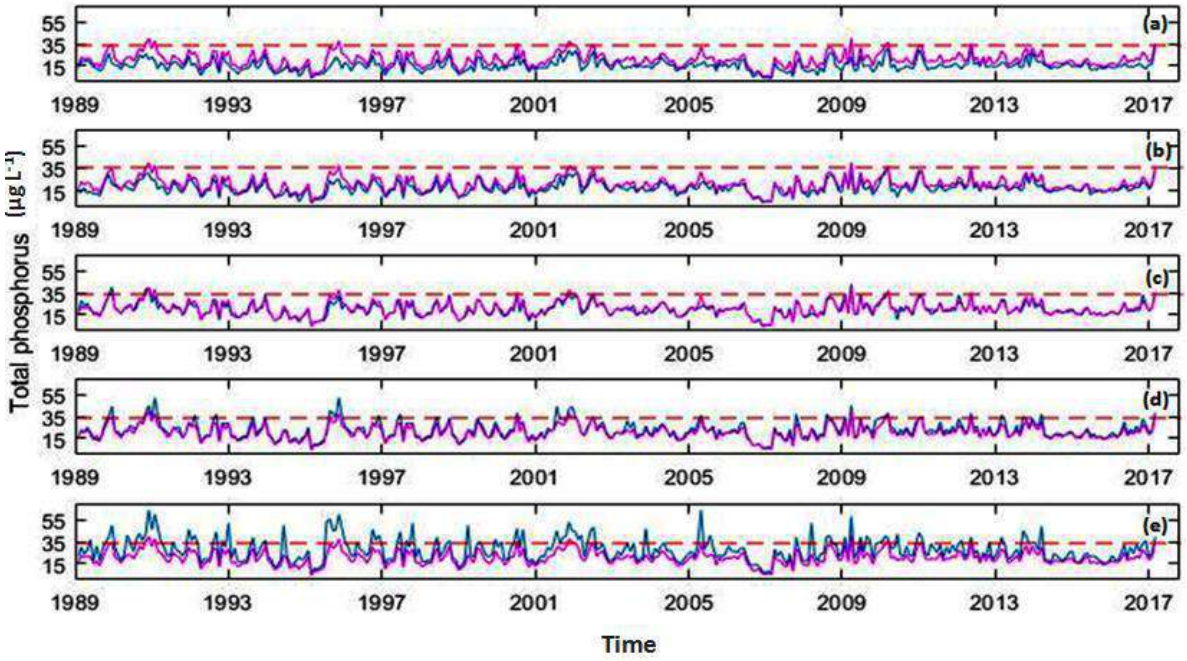



**Figure 20. Temporal variability of the total phosphorus concentrations in the North Lagoon of Tunis: a) Station 1; b) Station 2; c) Station 3; d) Station 4; e) Station 5; blue line: total phosphorus concentrations at the Station - purple line: lagoon mean concentration over the sampling Stations - Red line: OECD threshold.**

As found in most Mediterranean lagoons, TP in the North Lagoon of Tunis varied among the seasons with concentrations low in spring but high in autumn and winter (rainy season), which could be attributed to internal phosphorous loading (Figure 21).

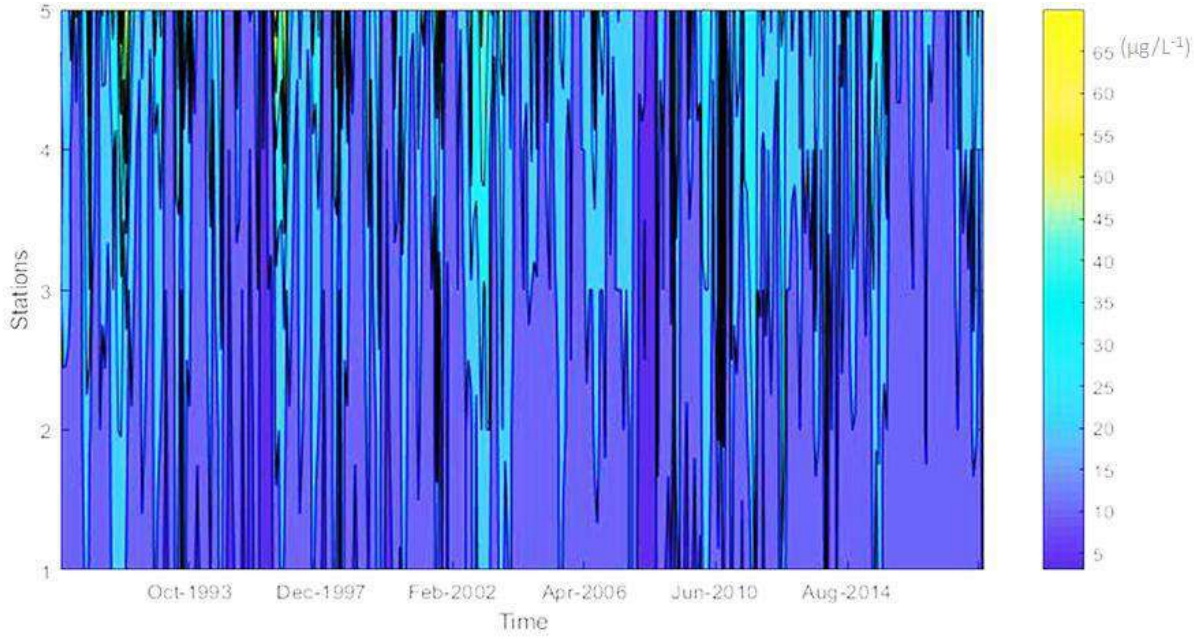



**Figure 21. Seasonal variation of total phosphorus concentrations in the North Lagoon of Tunis.**

### 4.1.5 Total nitrogen

The spatio-temporal variation of the total nitrogen (TN) in the North Lagoon of Tunis is presented in figure 22. Clearly, Stations 4 and 5, located at the south part of the lagoon are more loaded with TN than the north part of the ecosystem. TN exhibits the same spatial gradient than TP (Figure 20). This is explained by the water circulation system installed in the lagoon during the restoration project.

The higher concentration of total nitrogen recorded at Station 5, in comparison with the other Stations, is in agreement with the high level of all the nutrients measured in this area of the lagoon. These waters originate from a location (Station 4) that was not impacted by the dredging work carried out during the restoration project, and are distinguished by a nutrient-rich bottom. Thus, the relatively high concentrations in total nitrogen are in good agreement with the relatively high TP concentrations shown above in Figure 20.

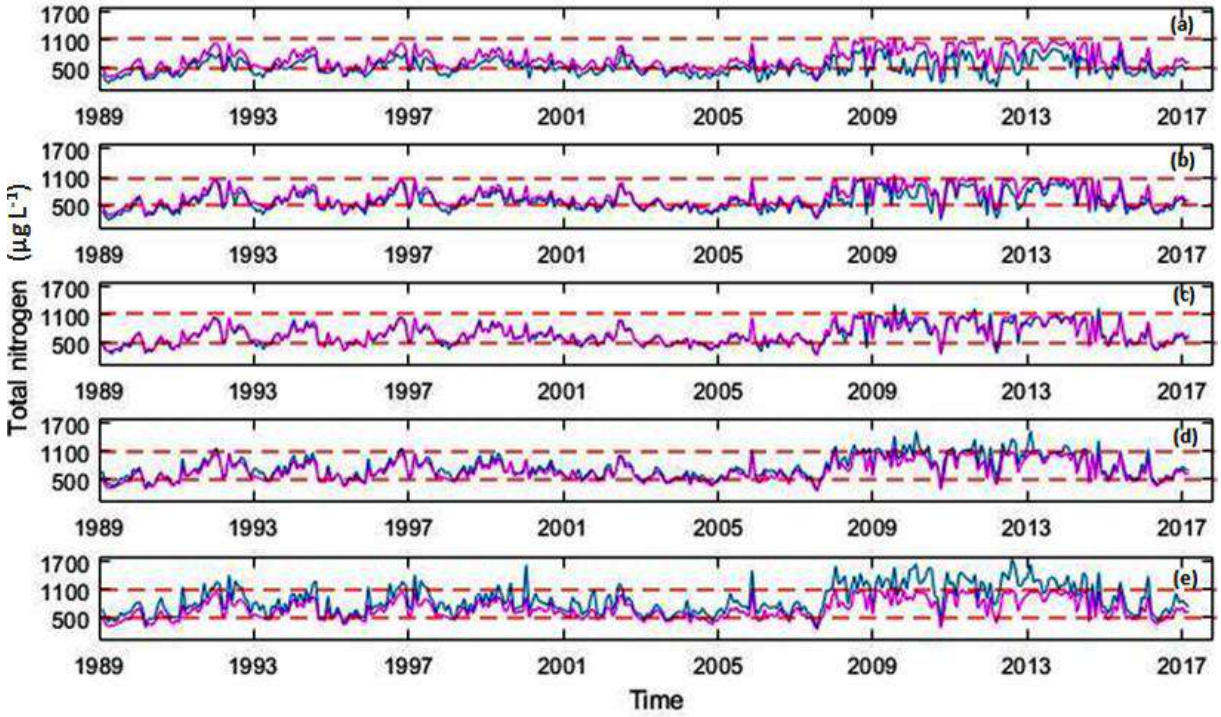



**Figure 22. Temporal variability of the total nitrogen concentrations in the North Lagoon of Tunis: a) Station 1; b) Station 2; c) Station 3; d) Station 4; e) Station 5; ; blue line: total nitrogen concentrations at the Station - purple line: lagoon mean concentration over the sampling Stations - Red lines: Wetzel thresholds.**

Nutrient concentrations (TP and TN) exhibit a common seasonal variation, being relatively high in autumn and winter and low in summer (Figure 23).

The high nutrient concentrations seen in the North Lagoon of Tunis during the winter can be attributed to external nutrient imports from the Gulf of Tunis, as well as phytoplankton growth.

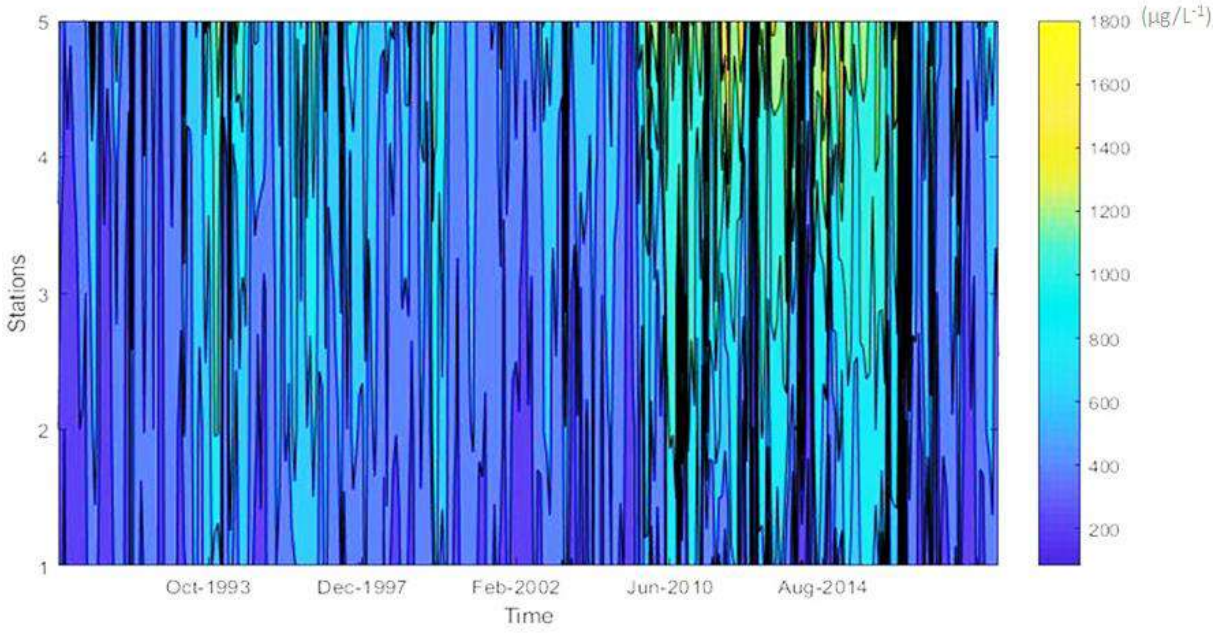



**Figure 23. Seasonal variation of total nitrogen concentrations in the North Lagoon of Tunis.**

### 4.1.6 Chlorophyll-*a*

The most essential pigment in aerobic photosynthetic organisms is chlorophyll-*a* (Chl-*a*). Indeed, it is measured to quantify the quantity of phytoplankton biomass in the water, and hence the degree of eutrophication (Tian *et al*., 2017).

The lowest Chl-*a* measurements were recorded between 2000 to 2009 and the highest Chl-*a* concentration were recorded during the first and last decades of our time series (Figure 24).

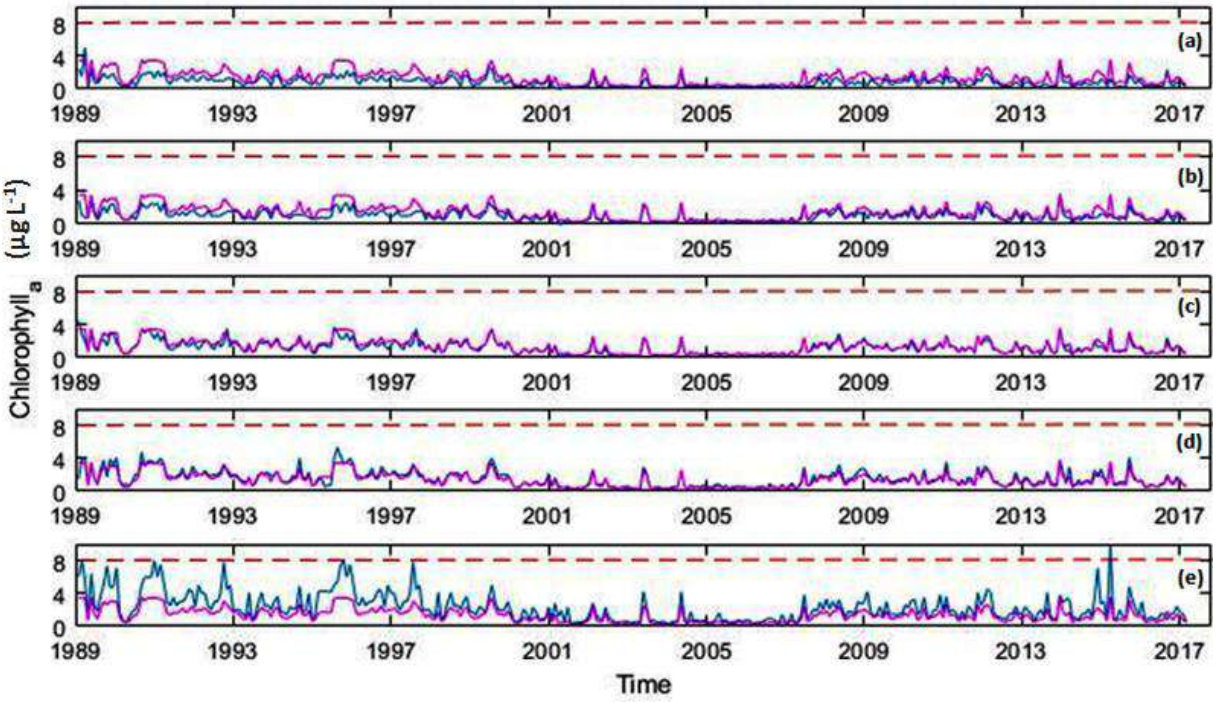



**Figure 24. Temporal variability of Chl-*a* in the North Lagoon of Tunis: a) Station 1; b) Station2; c) Station 3; d) Station 4; e) Station 5; blue line: Chl-*a* concentrations at the Station - purple line: lagoon mean concentration over the sampling Stations - Red line: OECD threshold.**

The Chl-*a* concentrations were high at the south part of the lagoon (Station 5; clearly exceeds the lagoon average), the values reached at Station 5, 10 µg $L^{-1}$ in 2015. This variation is mainly due to the water circulation system installed in the lagoon during the restoration work. In fact, the south part of the lagoon is located at the exit of the waters from the lagoon to the sea, via the Gulf of Tunis. In general, the water is loaded at this area.

The concentrations of Chl-*a* followed a seasonal trend (Figure 25), with concentrations being greater in the spring (13 - 16 °C). During the summer through winter (19 - 28 °C), Chl-*a* fluctuations were minimal and reasonably constant.

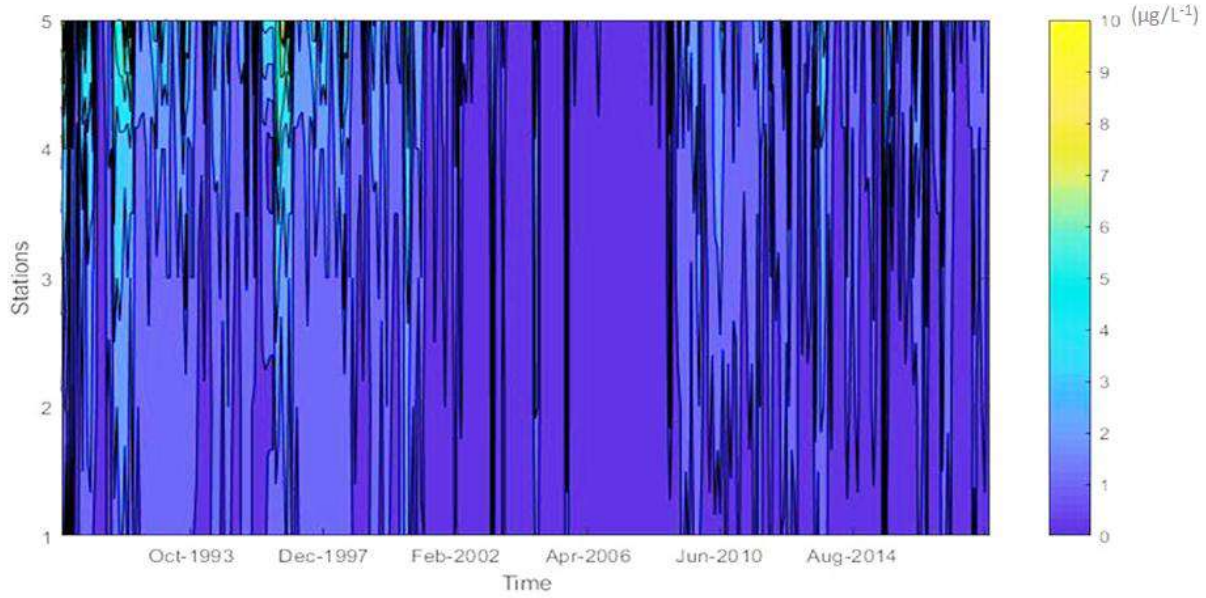



**Figure 25. Seasonal variation of Chlorophyll-*a* in the North Lagoon of Tunis.**

### 4.1.7 Dissolved oxygen

Important biological processes are associated with oxygen distribution, such as photosynthesis, respiration, and decomposition (Srichandan *et al*., 2015). Usually in coastal lagoons the occurrence of algal blooms is followed by eventual rate's oxygen decrease in the water column. This allows us to associate the quantitative variability of the phytoplankton as descriptor of environmental stress (Domingos *et al*., 2012).

Before the restoration project in the North Lagoon of Tunis, the dissolved oxygen levels were characterized by very remarkable fluctuations, with many periods of anoxia generating high mortalities of fish and organisms (Zaouali, 1974). Since the restoration of the ecosystem, the environment is much more oxygenated (Shili, 1995).

The level of dissolved oxygen (DO) ranged from 4.2 to 8 mg $L^{-1}$ and 8.7 to 11.5 mg $L^{-1}$ in summer and winter seasons, respectively (Figure 26). In reality, dissolved oxygen production is inversely related to temperature. Low temperatures reduce oxygen diffusion in the atmosphere while increasing its solubility in water (Belkhir et Haj Ali, 1981). In addition, significant algal death has been observed in the North Lagoon of Tunis due to high

temperatures. The decomposition of this dead biomass needs DO consumption, which makes its level lower in hot episodes. Figure 26 shows that the temporal (seasonal and interannual) variations of dissolved oxygen are very similar at all five Stations in the lagoon. This figure further illustrates that the lagoon remains well oxygenated from the inlet to the outlet gates. Thus, suggesting that the dissolved oxygen air-sea exchanges are efficient and that the lagoon ecological state remains healthy.

The variations of the dissolved oxygen seem to be more important at Station 4. This can be due to its shallowness, which facilitates the exchanges between the atmosphere and the water.

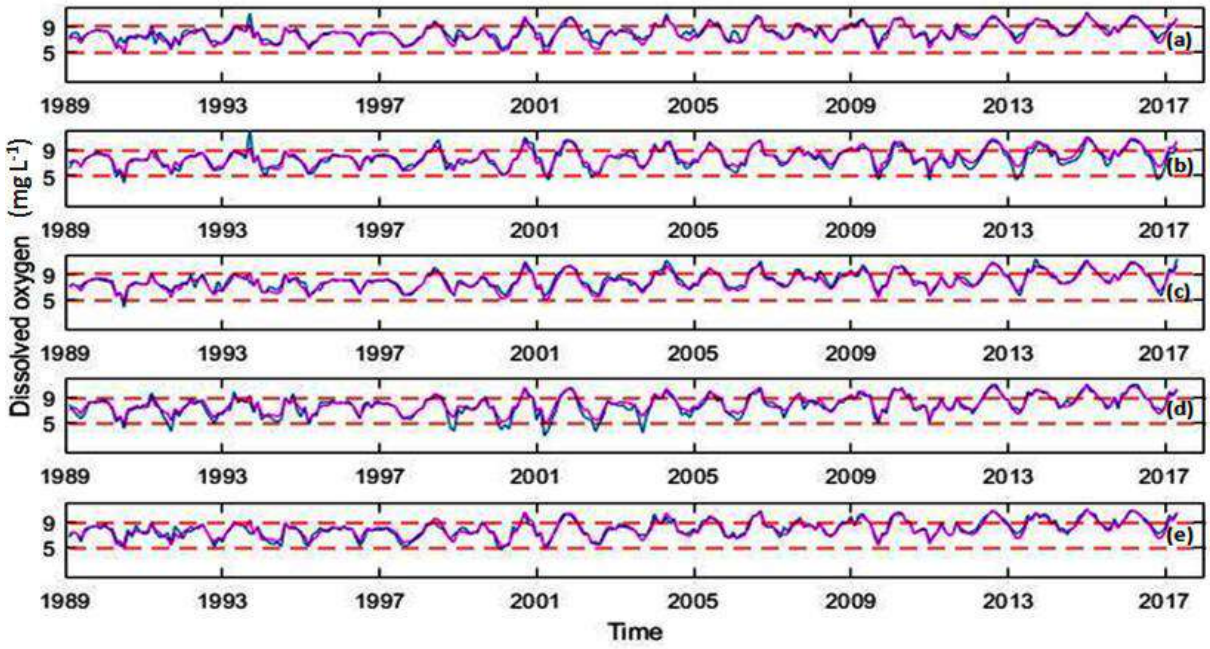



**Figure 26. Temporal variability of the dissolved oxygen in the North Lagoon of Tunis: a) Station 1; b) Station2; c) Station 3; d) Station 4; e) Station 5; blue line: dissolved oxygen concentrations at the Station - purple line: lagoon mean concentration over the sampling Stations - Red lines: EU and WHO thresholds.**

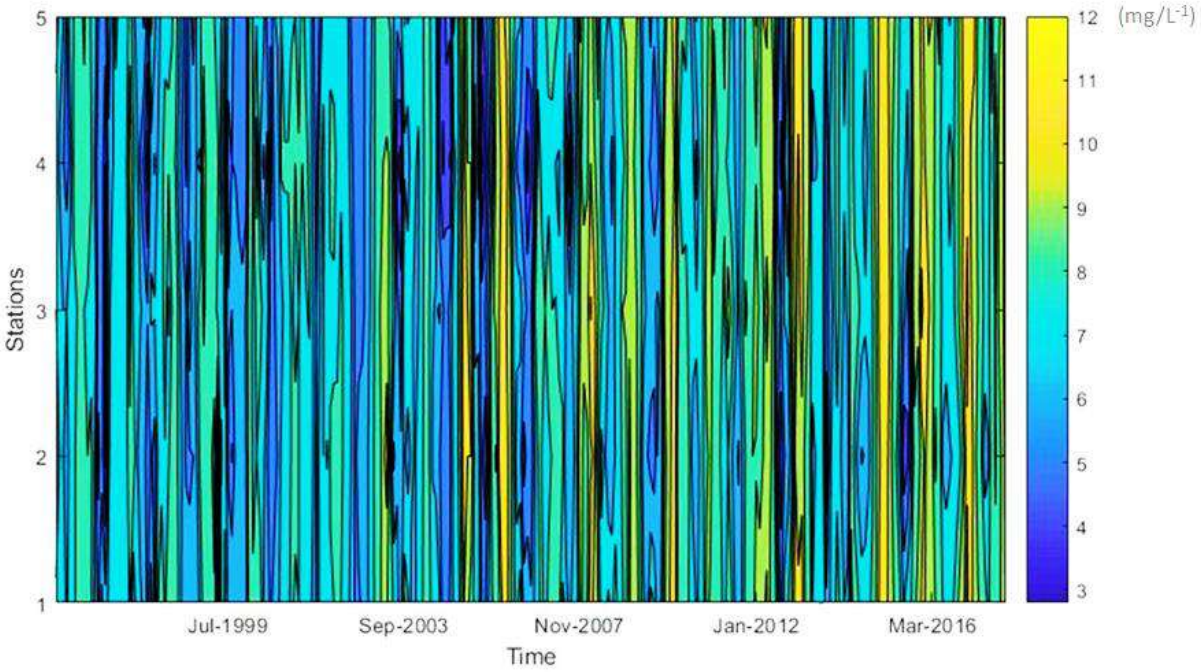



**Figure 27. Seasonal variation of dissolved oxygen in the North Lagoon of Tunis.**

### 4.1.8 Salinity

Salinity is one of the most important variables influencing the dispersion of living organisms. Indeed, its variations can significantly alter their physiological activities as well as their ecological tolerances (Kinné, 1958). Salinity occurs as a key component in studies of the physicochemical properties of water, which is a result of numerous processes (natural or manmade).

Water salinity level in the North Lagoon of Tunis ranges between 30 to 46 with a mean of about 38 (Figure 28). It is worth mentioning that before the restoration of the North Lagoon of Tunis, salinity was about 50 (Belkhir, 1980).

High salinity values are recorded at Station 4 and 5 (south part of the lagoon) and the lowest values are recorded at Station 1 and 2 (north part of the lagoon). This gradient is similar to that of the nutrients and Chl-*a*, which may be explained by the system of water circulation in the lagoon established during the restoration program. It is clear that salinity at all Stations, in addition to the lagoon mean values are below the amount recorded before the restoration of the North Lagoon of Tunis.

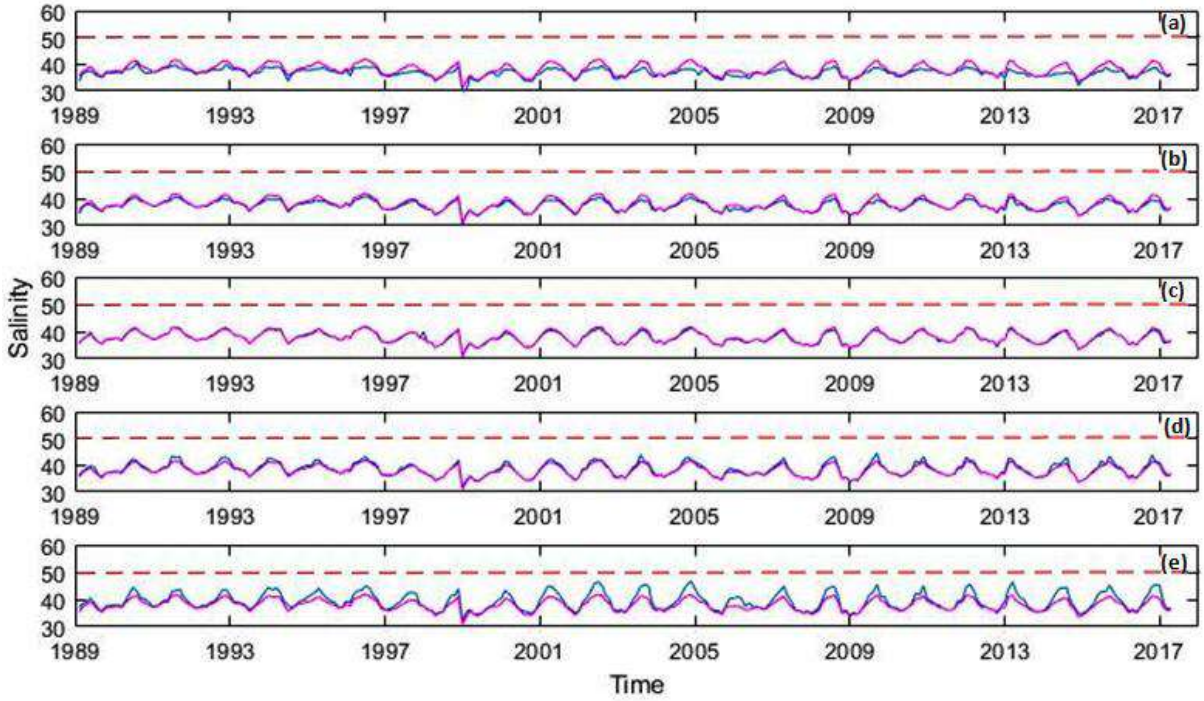



**Figure 28. Temporal variability of the salinity in the North Lagoon of Tunis: a) Station 1; b) Station 2; c) Station 3; d) Station 4; e) Station 5; blue line: salinity values at the Station - purple line: lagoon mean concentration over the sampling Stations - Red line: salinity level before the lagoon restoration.**

Salinity was low during the winter season and gradually increased as the season advanced towards summer. On the one hand, this can be explained by water evaporation during high-temperature events throughout the summer season, as well as a decrease in precipitation and water runoff. On the other hand, high dilution during the rainy season, might lead to a drop in salt concentration in the lagoon (Figure 29).

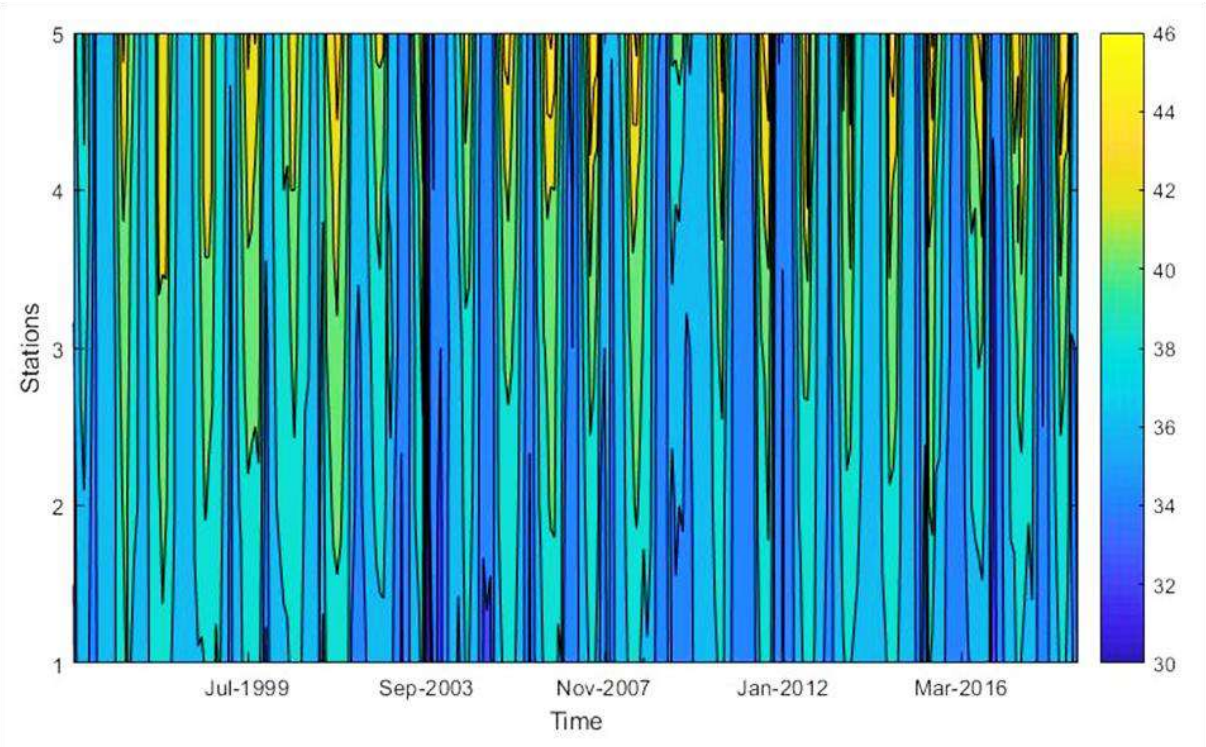



**Figure 29. Seasonal variation of the salinity in the North Lagoon of Tunis.**

## 4.2. Discussion

The mean temperature of the North Lagoon of Tunis (19 °C) is comparable to the mean values of about 19.22 °C and 19.27 °C in Sardinian ecosystems, Calish lagoon and Santa Guista lagoon in Italy, as reported by Satta *et al*., 2020.

In addition, the mean temperature of the North Lagoon of Tunis is similar to that of other Tunisian ecosystems, such as Bizerte Lagoon (19.69°C) and Ghar el Melh Lagoon (19.35°C), as reported by Béjaoui *et al*. 2016., 2018.

Before the restoration project, in the North Lagoon of Tunis, Ben Charrada (1992) recorded pH values that ranged from 9.2 to 6.4 and this may be due to significant photosynthetic activity. After the restoration, Shili (1995) recorded values that ranged between 8.2 and 8.4.

Similarly to the North Lagoon of Tunis which pH varies between 8.7 and 7.2, Ouijid *et al*., 2020, found that the pH of another Mediterranean lagoon (Nador Lagoon) along the Morocco's Mediterranean coast fluctuated from 7.73 to 8.2.

The pH of the surface water in the North Lagoon of Tunis showed a slight decline since 2010 and an alkaline tendency during all the seasons, in the five Stations. In natural waters, generally, pH values vary between 6 and 8.5 (Chapman *et al.*, 1996). A pH range of 6.5 to 8.5 is acceptable for aquatic biota according to the American Public Health Association (APHA, 1999). As a result, we may conclude that the North Lagoon of Tunis is in good agreement with these limits.

According to the Organization for Economic Development and Cooperation (OECD, 1982), the hyper eutrophic lakes generally showed the maximum transparency values at ≤1.5 m and minimum transparency ≤0.7 m. This indicates that the North lagoon of Tunis is a highly productive ecosystem with areas affected by eutrophication.

As compared with other Mediterranean lagoons, the average Secchi depth of the North Lagoon of Tunis (2 m) is deeper than Calish lagoon (1.2 m), Santa Guista lagoon (1 m) as reported by Satta *et al*. (2020). We can also mention Or lagoon (1 m) as specified by David *et al*. (2019). Also, the North Lagoon of Tunis is shallower than other Mediterranean lagoons like Thau lagoon, which has an average depth of 4 m (Delorez *et al*., 2020).

The TP values in the lagoon vary between 4 – 65 µg $L^{-1}$, which is greater than the ecosystems at the Palvasian lagoon complex (varies between 0.9 - 12 µg $L^{-1}$) reported by Leruste *et al*., 2015. According to the Organization for Economic Development and Cooperation (OECD, 1982) for trophic state classification, the limit of TP to define eutrophic ecosystem is 35.0 µg $L^{-1}$. Considering this, we must point out, that the south part of the lagoon, especially, Stations 4 and 5 are eutrophic areas in the ecosystem.

According to Wetzel (2001), total nitrogen concentrations ranging from 500 and 1100 µg $L^{-1}$ are considered to be thresholds for eutrophication in aquatic habitats. Figure 22 clearly shows that Stations 4 and 5, located in the south area of the lagoon, exceed those limits especially since 2007.

According to OECD (1982), classification of trophic state, 8.00 µg $L^{-1}$ of Chl-*a* concentration is the threshold for eutrophication, and in the present study; Station 5, were found to have Chl-*a* concentrations beyond the prescribed limit. As compared with other lagoons, the concentrations of Chl-*a* in the North Lagoon of Tunis (0.22 - 10 µg $L^{-1}$) was lower than some Tunisian lagoons, such as, Ghar el Melh lagoon (1.5 to 9.7 µg $L^{-1}$), Bizerte lagoon (the average value is about 7.20 µg $L^{-1}$), as recorded by Béjaoui *et al*., 2018; 2016, respectively.

The concentration range of the DO in the North Lagoon of Tunis (4.2 -11.5 mg $L^{-1}$), is almost similar to most other Mediterranean lagoons. In fact, dissolved oxygen in Nador lagoon located in Morocco, varied between 3.84 and 13.8 mg $L^{-1}$ (Ouijidi *et al*., 2020). According to the World Health Organization (WHO, 1996) and European Union (EU, 1998), the standard for DO value for fisheries and aquatic life is between 5.0 and 9.0 mg $L^{-1}$. Concentrations below 4.0 mg $L^{-1}$ adversely affect aquatic life. The lagoon approximately fits in the range of WHO (1996) and EU (1998) guidelines, but slightly exceeds the maximum permissible values.

In comparison with other Mediterranean lagoons, salinity in the North Lagoon of Tunis is higher than in Thau lagoon, where salinity ranges between 29.6 and 40.8, according to Derolez *et al*. (2020), and Calish lagoon (mean salinity of about 17.73) and Santa Guista lagoon (mean salinity of about 32.13) reported by Satta *et al*. (2020). We can also mention Or lagoon where salinity values are between 2 to 35 according to David *et al*. (2019).

The salinity in the North lagoon of Tunis can be relatively similar with Tunisian Mediterranean lagoons. In fact, in Ghar Melh lagoon, salinity ranges between 26.6 and 51.2

(Béjaoui *et al*., 2018). Also, in Bizerte lagoon, salinity showed fluctuations between 34.2 to 38.4 according to Béjaoui *et al*. (2016).

## 4.3. Analysis of variance (ANOVA)

ANOVA revealed no significant difference among the sampling Stations with p-values > 0.05 in the lagoon for any of the physico-chemical variables or Chl-*a* (Table 3). Therefore, all data were grouped by month to reconstruct the monthly dynamics of each variable in the North Lagoon of Tunis.

**Table 3. ANOVA results (p-values)**

| | **Chl-*a*** | **SD** | **DO** | **TP** | **TN** | **Temperature** | **Salinity** | **pH** |
|---|---|---|---|---|---|---|---|---|
| **Station 1** | 0.09 | 0.51 | 0.44 | 0.64 | 0.97 | 0.56 | 0.82 | 0.06 |
| **Station 2** | 0.29 | 0.32 | 0.58 | 0.79 | 0.82 | 0.54 | 0.69 | 0.12 |
| **Station 3** | 0.25 | 0.28 | 0.45 | 0.56 | 0.64 | 0.45 | 0.52 | 0.09 |
| **Station 4** | 0.36 | 0.08 | 0.52 | 0.49 | 0.73 | 0.53 | 0.72 | 0.23 |
| **Station 5** | 0.06 | 0.4 | 0.47 | 0.38 | 0.69 | 0.61 | 0.91 | 0.31 |

## 5. Conclusion

This first chapter gives an overview of the data and the spatio-temporal distribution of the physico-chemical characteristics of the lagoon, as well as their seasonal changes.

Sampling cruise, analyzes in the laboratory, data provided by Al Buhaira Invest company, as well as the bibliographic study enabled us to reconstruct a time series of 7 physico-chemical variables and one biological (Secchi depth, total nitrogen, total phosphorus, pH, salinity, dissolved oxygen, water temperature and chlorophyll-*a*) for approximately three decades (January 1989 - April 2018). The spatio-temporal and seasonal representations of these variables enabled us to conclude that:

The water temperature gradually increases as the ambient air temperature increased. But we have not reported a sudden increase, which confirms the absence of thermal pollution in the study area. It should be noted that prior to the restoration project, the studied lagoon was subjected to thermal pollution produced by discharges from the Tunisian Electricity and Gas Company.

The salinity depends on the position of the sampling location; the northern part of the lagoon is characterized by a lower salinity concentration than the southern part. In fact, the marine water from the Gulf of Tunis comes into the lagoon at the north part. The strategy of water circulation allows its loading with nutrients and suspended matter during its residence time in the lagoon, which contribute to the increased salinity in the southern part. In addition, in the southern part, the evaporation is more intense (shallower depth), because, while realizing the restoration project, the dredging works mainly concerned the north part of the lagoon.

The salinity content also depends on the season due to the influence of the rainfall especially at the Stations near the areas where the rainwater is discharged (Stations 1, 2 and 3).

The recorded pH values do not exceed the standards according to APHA (1999) and which are of a basic nature (pH>7).

Dissolved oxygen (DO) is an essential parameter in the maintenance of aquatic life, such as the process of degradation of organic matter and photosynthesis. In the North Lagoon of Tunis, DO contents were higher in the northern side, where the marine waters from the Gulf of Tunis comes in, rather than the southern side where the waters are already loaded, which reduce the DO content.

Some relatively high values of Chl-*a* concentrations, as well as total nitrogen and total phosphorus, may indicate a state of eutrophication in certain areas of the lagoon, particularly in the southern part.

Among the environmental and biological data, dissolved oxygen, total phosphorus, salinity and Chl-*a* values were the most affected by the seasonal variations.

The highest values were recorded for salinity, total phosphorus and total nitrogen during the summer (or dry) season. In spring, Chl-*a* followed by DO showed high values.

Most of the parameters exhibited a significant spatial variation. They revealed that the coastal water was significantly influenced by freshwater input from the gulf of Tunis. In general, the physico-chemical water quality agrees fairly well with the marked improvement in the eutrophication state of the North Lagoon of Tunis.

According to our data, the iterannual trends of the Chl-*a* can be divided in 5 periods. Between January 1989 and May 1996, the Chl-*a* concentrations are approximately similar. There has been a relatively large increase between June 1996 and December 1996. The period between January 1997 and February 2001 is similar to the first one. In Mars 2001 to September 2008, a significant decrease in the Chl-*a* concentrations has been noticed which shows the good functioning of the ecosystem and the gradual decrease in the lagoon trophic level. Since 2008, an increase in the Chl-a content is noticed but it's less significant than the previous rise in Chl-*a* concentrations between 89-96.

The North Lagoon of Tunis is now a totally artificialized environment as a result of human intervention, and ecological follow-up is required to ensure the proper ecological functioning of this ecosystem located in the heart of the metropolitan zone.

Knowing that, predictive modeling seems to be an appropriate technique to work with in the North Lagoon of Tunis, in the purpose to predict and anticipate any deterioration or pollution phenomenon in this natural ecosystem.

In order to determine the most appropriate algorithm, pre-processing the data like it is done in this chapter, is a crucial step to capture any hidden pattern in the distribution of the environmental variables like spatial, temporal, seasonal, linear, nonlinear, etc.

Zhang and Qi (2005) created a predictive model that takes seasonality into account, emphasizing the need of appropriate data pre-processing to account for seasonal or trend fluctuation.

Our data appears to have nonlinear and seasonal characteristics. Indeed, before doing any complex modeling we need to take those specificities into account. Then, the most adequate technique chosen will be one of the tools used to monitor the ecological status of the ecosystem.

# Chapter II: Machine Learning modeling techniques for forecasting the trophic level in a restored South Mediterranean lagoon using Chlorophyll-*a*

## 1. Introduction

Chlorophyll-*a* is a commonly used environmental indicator of algal biomass and lake eutrophication (Latif *et al*., 2003). High Chl-*a* concentrations may be interpreted as the presence of algae blooms in coastal waters, which are one of the most serious environmental challenges, due to their negative economic and environmental consequences on water bodies (Pearson *et al*., 2001).

As stated earlier in this study, Chl-*a* *in-situ* sampling and measurement programs require routine monitoring and laboratory analysis. Because of various constraints, these programs have limited environmental management capabilities to effectively track and respond to eutrophication occurrences, such as (i) field monitoring budget, (ii) availability of personnel and equipment, (iii) field safety concerns and (iv) lengthy time intervals between data collection, reporting, and public notification. Therefore, in order to reduce the cost and time required for aquatic environmental *in-situ* monitoring and laboratory analyses, a predictive modeling approach that provides the Chl-*a* values automatically is essential to prevent or mitigate the occurrence of eutrophication and, ultimately, to minimize the potential adverse effects on water bodies (Oh *et al*., 2007).

Data-driven Machine Learning (ML) models are popular and widely used for modeling complicated natural processes, particularly in predictive modeling. They can be useful in modeling and predicting eutrophication events in any natural environment (Nayak *et al*., 2005; Wu *et al*., 2011; He *et al*., 2014).

The factors influencing Chl-*a* concentrations in water bodies are diverse and complicated (Jimeno-Sáez *et al*., 2020). Various mathematical approaches have been used in current literature to forecast Chl-*a* based on regression analysis. However, these traditional data processing methods generally employ a linear connection to simplify complex situations, resulting in poor results since they are inefficient in dealing with complicated non-linear relationships between the variables involved (Su *et al*., 2015).

Because they are ideally adapted for forecasting non-linear and complicated functions, ML algorithms have been proven to be more efficient than traditional data processing approaches in monitoring water quality (Abba *et al*., 2017). Previous studies have shown that ML outperforms traditional techniques to modeling water quality factors (Juntunen *et al*., 2012; Charulatha *et al*., 2017).

Artificial neural networks (ANN) algorithms are used intensively among ML techniques. It imitates human learning processes by network preparation and calibration. This skill makes ANN useful tools for analyzing dynamic situations that are difficult to explain using traditional methods (Daliakopoulos *et al*., 2005; Samarasinghe, 2007). The ability to capture system dynamics and nonlinearities makes ANN particularly suited for the investigation of natural environments, which typically have distinctive spatial–temporal heterogeneity (ASCE, 2000).

The ANN algorithms were also applied to Chl-*a* dynamics since it is one of the variables that represents algae biomass and has been recognized as one of the proactive early-warning techniques to preventing the occurrence of some algal blooms. Li *et al*. (2017) and Yi *et al*. (2018) applied different types of artificial neural networks to estimate the concentration of Chl-*a* in 27 lakes in China and in one Korean river, respectively. Tian *et al*. (2017) used an ANN to predict Chl-*a* concentrations to an estuary reservoir in East China.

Back Propagation networks, Radial Basis Function networks, and other categories exist in artificial neural networks. Back propagation is a learning method that is widely employed in ANN applications. In this study, a nonlinear autoregressive with external inputs (NARX) neural network, a contemporary ANN method, was developed. NARX is a dynamic neural network that belongs to the non-linear back propagation class (Markova, 2019).

In terms of time consuming and high cost, reducing the number of parameters to be measured is very important. For this reason, it is of a big interest to select specific variables that are most related to Chl-*a* concentrations. The Random Forest (RF) approach, developed by ML, is an effective method for accomplishing this task. RF has been applied in many studies. Béjaoui *et al*. (2016) investigated with the RF model the most important predictor variables for Chl-*a* variation in the lagoon of Bizerte situated in the north of Tunisia. In another research study, Béjaoui *et al.* (2018) used the RF model to investigate the dynamics of the plankton in Ghar Melh lagoon located in the Tunisian Mediterranean coast.

As well known, an early-warning proactive approach of the Ch-*a* content is essential to prevent the occurrence of eutrophication episodes, especially in sensitive ecosystems. For this reason, we used NARX to perform a one-step ahead forecasting of the Chl-*a* concentration in the North Lagoon of Tunis.

Given the superiority of the ML algorithms, this study has been conducted to achieve the following objectives: (1) to select the specific variables that are the most related to Chl-*a* concentrations in the North Lagoon of Tunis and Station 5, using, especially, RF model. To do so, different variables combinations were tested, (2) to develop an ANN network to estimate and forecast one step ahead of Chl-*a* concentrations based on NARX neural network and to (3) validate the performance of the model in the lagoon and at Station 5.

Several studies have been conducted on the eutrophication process and water quality indicators of the North Lagoon of Tunis (Ben Charrada, 1992; Rezgui *et al*., 2008; Trabelsi-

Bahri *et al*., 2013). However, to the very best of our knowledge, there is no previous research using ML models to predict water quality parameters in this lagoon, specifically, Chl-*a* concentrations.

## 2. Machine Learning (ML)

Building that precise mathematical formalisms is difficult for ecologists to consider and explain, and even more hard to do so in a programming language (Kompare *et al*., 1994). If these tasks have been completed, the computational modeling traps, data lack and inconsistency, difficulties with the correct modeling parameters, and so on might be challenging. We would all welcome a computer program that could understand our common descriptions of ecosystems and convert them into a machine code (Kompare *et al*., 1994).

Recently, data-driven Machine Learning techniques, also known as Artificial Intelligence, have gotten a lot of attention because of their ability to solve complicated multivariate nonlinear problems (Nyshadham *et al*., 2019). This is done, by developing mathematical models to describe relationships between inputs (influence factors) and outputs (Hayajneh *et al*., 2009).

Performing Machine Learning means creating a model. The best way to define Machine learning approaches is to state that ML teaches computers to do what comes naturally to humans: learn from experience (Williams and Poff, 2006).

Machine learning algorithms are employed in a variety of situations when it is difficult or impossible to build an algorithm that properly describes a complex natural process (Kompare *et al*., 1994). That is, machine learning algorithms employ computational methods to "learn" information directly from data, rather than depending on a preset equation as a model (Williams and Poff, 2006). It is to be mentioned that machine learning allows the use of vast

amounts of data (Kompare *et al*., 1994). As the number of samples available for learning increases, the algorithms performance improves adaptively [1].

Machine Learning (ML) has made significant development in recent years, and it has been widely applied in many new applications where data can be collected and processed locally (Gohel *et al*., 2019). These data may be used to train machine learning models, which can then be used to make predictions, and helping in management decisions making in a variety of applications (Wu *et al*., 2011).

Machine learning employs two techniques (Figure 30): supervised learning, which creates a model of known input and output data in order to predict future outputs, and unsupervised learning, which seeks out hidden patterns or intrinsic structures in data [1].

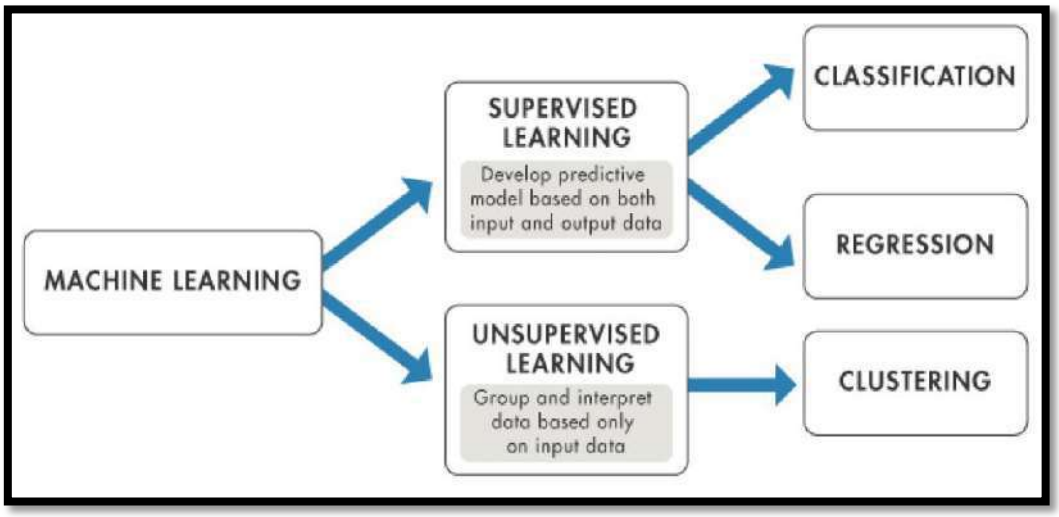



**Figure 30. Representation of the machine learning techniques [1].**

There are several of supervised and unsupervised machine learning algorithms, each has its own method to learning (Figure 31). Choosing the best technique depends on the type of data considered in input and output and the type of task or problem to be solved, which can be overwhelming. There is no optimum approach, nor is there a one-size-fits-all solution. The search for the best algorithm is largely based on trial and error, and even highly competent computer scientists cannot predict whether an algorithm will work without trying it [1].

Highly adaptable models have a tendency to over-fit data by simulating small changes that may or may not represent noise. Simple models are easy to comprehend, but their accuracy is frequently poorer. As a result, selecting the best algorithm leads to balancing benefits such as model speed, accuracy, and complexity. ML is built around trial and error. If one method or algorithm fails, the only option is to try another [1].

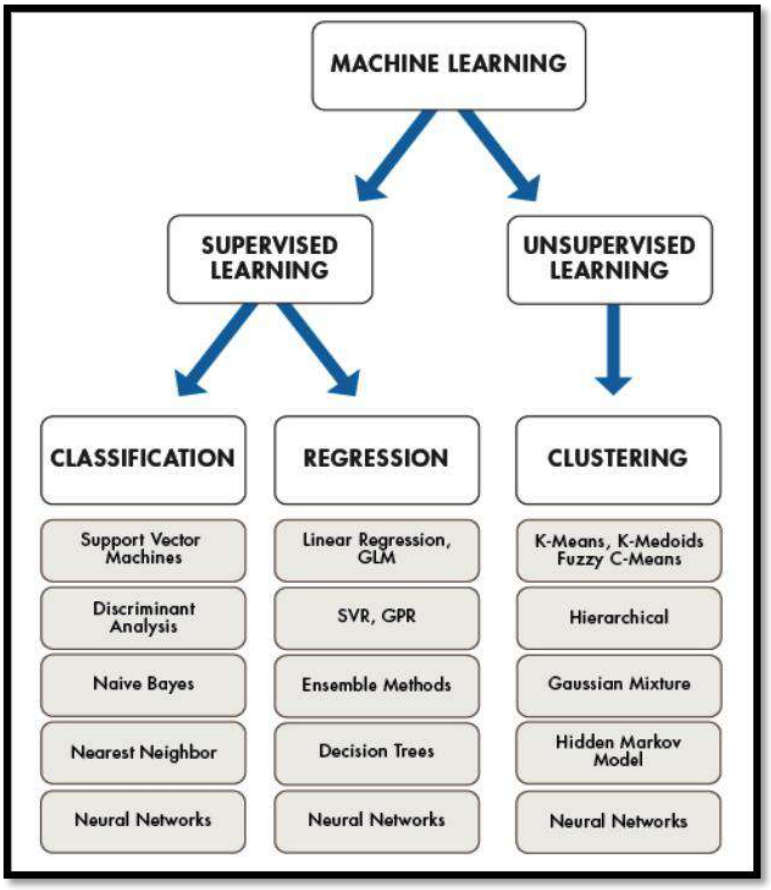



**Figure 31. Machine Learning algorithms [1].**

Netflix's inaugural "Netflix Prize" competition was launched in 2006 to discover a program that could better predict customers' interests and enhance the accuracy of its current Cinematch movie recommendation system by at least 10% using machine learning techniques [2].

In 2010 The Wall Street Journal wrote about the firm Rebellion Research and their use of machine learning to predict the financial crisis (Patterson, 2010).

In 2012, co-founder of Sun Microsystems, Vinod Khosla, predicted that 80% of medical doctors' jobs would be lost in the next two decades to automated machine learning medical diagnostic software [3].

Social Medias are the most widely used interfaces on the planet. According to estimates, there are over two billion users. Instagram, YouTube, Twitter, Tik Tok, Facebook, and other social media platforms are employing powerful machine learning algorithms to propose content and advertisements to viewers based on predictive modeling.

# 3. Methodology

## 3.1. Random Forest (RF)

ML algorithms are generally performed using a collection of predictor variables (input variables) and one or more target variables (output variables) expressed as a continuous value (Kohavi and Jhon, 1997), where the fundamental aim of predictive modeling is to maximize accuracy (Motoda and Liu, 2002). To estimate a parameter of water quality we can use all available predictor variables or select a smaller number of them. This might result in the model having too few or too many inputs, both of which are undesirable (Maier *et al.*, 2010). This study performed a predictor variable selection stage to minimize redundant data in order to address this issue. The objective of reducing the number of predictor variables in ML is to speed up the learning algorithm process, increasing prediction accuracy and comprehensibility of learning findings (Motoda and Liu, 2002). It is known that Chl-*a* concentrations are influenced by a variety of factors. The Random Forest model was used in this work to find the most relevant predictor variables for Chl-*a* concentrations.

RF modeling is a relatively recent ML method that is built on multiple decision trees and trained on a set of input predictor variables to predict the output variable accurately (Breiman, 2001; Strobl *et al*., 2006; Strobl *et al.*, 2008).

The RF method has several advantages. First, there is no assumption of a probability distribution for predictor variables. Second, it is capable of dealing with a huge number of variables and picking the most helpful ones among them (Mulia *et al*., 2013; Park *et al*., 2015). Third, RF predictions are extremely reliable since they are based on the average of several basic models, avoiding the over-fitting issue typical of many non-linear regression approaches (Phillips *et al*., 2008; Huang *et al*., 2015).

Fourth, because each tree is created on a random subset of the original data, there is no need for a separate independent dataset or cross-validation technique to assess the model's prediction ability (Were *et al*., 2015). Finally, the RF method is appropriate for natural ecosystems with a high number of physicochemical and biological variables with complicated interactions.

The RF model has this interesting aspect of providing a quantitative assessment of the relevance of the predictor variables in the final outcome, which can be useful in selecting the most essential ones.

The method used to evaluate the ranking of the most important predictor variables from the RF model is the out-of-bag (OOB) technique by permutation; a technique that measures how influential the predictor variables in the model are at predicting the response variable (Chl-*a*). The effect of the predictor variable increases with the value of this measure.

If a predictor variable influences the prediction, then the permutation of its values should have an impact on the model error. If a predictor variable is not influential, the permutation of its values should have little to no effect on the model error (Mitchell, 2011). It consists of calculating the gain in the mean square error, which is computed by permuting OOB data: for each tree, the prediction error on the OOB portion of the data is recorded; the same is done by permuting each predictor variable (Mitchell, 2011). The differences between the two OOB

errors are then averaged over all trees and normalized by the standard deviation of the differences (Mitchell, 2011).

In Machine Learning (ML), a variety of measures are used to assess a model's prediction accuracy. The choice of accuracy metric depends on the ML task. In ML it is advisable to use as minimum mathematical measurements as possible to evaluate a model [4]. It helps in a fair and accurate assessment among all the models to be tested.

The mean of squared errors (MSE) and coefficient of correlation (R) were used to evaluate RF model performance.

MSE is a numerical measure of the model's error when it makes predictions for the target variable. It is sensitive to outliers, and should be used in conjunction with other metrics to evaluate a particular model, when the data studied is messy (Cutler *et al*., 2007). If the MSE is close to 0, it indicates a very close approximation to the actual values. The MSE is defined as:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

Where:

$y_i$ and $\hat{y}_i$ denote the modeled concentrations and the observed concentrations of Chl-*a*, respectively and $n$ is the amount of data in each data set.

Prediction accuracy R represents the degree of correlation between the prediction values and the observed values, and a high R value (close to 1) means the prediction is close to the observed value (Xu *et al*., 2019).

$$R = \left(1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}\right)^{\frac{1}{2}} \quad (2)$$

Where: $\bar{y}_i$ denotes the average of the observed values of Chl-*a*.

It is to be mentioned that the coefficient of determination ($R^2$) was calculated from R to contribute to the assessing and comparing the performance of the models. Thus, useful

information can be obtained concerning the relative importance of all variables and their capability of forecasting Chl-*a* concentrations.

The RF model was simulated twice. First, we considered only the seven physico-chemical predictors we had for predicting Chl-*a* concentrations. Second, we examined if there was any evidence of spatial or seasonal dependency among Chl-*a* predictors. In other words, we investigated whether the predictions of Chl-*a* concentrations might be improved by taking sampling Stations and seasons into account as observable factors. This was accomplished by introducing two additional categorical predictor variables, one for each of the five Stations and the four seasons. This was conceivable because RF models can handle both quantitative and qualitative predictor variables.

RF modeling was performed using the MATLAB software MATLAB® software (version 9.3.0.948333 (R2017b), The Mathworks, MA, USA).

## 3.2. Artificial Neural Networks (ANNs)

Among the different approaches, the artificial neural networks (ANNs), were widely applied in the last decades in the fields of bioinformatics (Dopazo *et al*., 1997), ecology (Lek *et al*., 1996; Lek and Guegan, 1999), and environmental engineering (Singh *et al*., 2009; Hill and Minsker, 2010; Sahoo *et al*., 2013). In fact, the good performance of ANNs in various ecological models was verified (Park *et al*., 2003; Song *et al*., 2013).

The human brain is extraordinarily sophisticated and, quite literally, the most powerful computer engine ever known. ANNs are computer neural networks that are inspired by biological neural networks. The ANN method's ultimate goal was to solve problems in the same way as a human brain can. However, over time, attention moved to performing specific tasks, leading to deviation from biology (Rajaee *et al*., 2019). Artificial neural networks have been used to perform a wide range of tasks. They can be used to

estimate functions that are usually unknown (medical diagnosis, computer vision, voice recognition, etc.) or to forecast future values of potentially noisy time series based on historical data. A typical ANN is made up of numerous components known as neurons (processing elements) and connections that connect them. Neurons with comparable characteristics are clustered together in a single layer. An ANN typically has three distinct layers (Figure 32), including the input, hidden, and output layers (Rajaee *et al*., 2019).

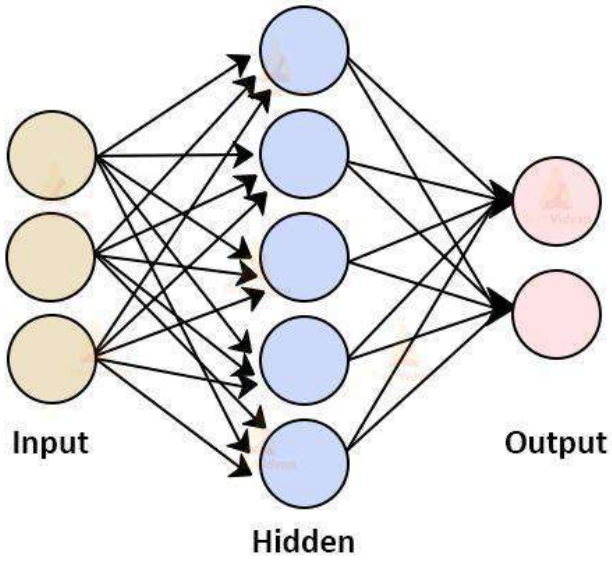



**Figure 32. Artificial Neural Network architecture [5].**

The network consists of connections, and a given neuron can have multiple input and output connections (Markova, 2019), each connection providing the output of one neuron as an input to the next neuron. Each connection is given a weight that changes as the learning process progresses and indicates its relative performance (Dogan *et al*., 2009).

The "weight" analysis computes the strength of the connections between the input factors and the output factors quantitatively and can be employed to explain the relations between the input variables and the output variable in ANNs (Garson, 1991).

In the hidden and output layers, each neuron passes its weighted input through a transfer (activation) function to produce a result. The activation function key feature is that it provides

a smooth, differentiable transition as input values change, in other words, a small change in the input produces a small change in the output (Rajaee *et al*., 2019). Then, the propagation function computes the input to a neuron from the outputs of its predecessor neurons and their connections as a weighted sum (Rajaee *et al*., 2019). A bias term can be added to the result of the propagation (Rajaee *et al*., 2019).

ANNs are trained with sample data, so that an input leads to a specific target output. Training means tuning the adjustable network parameters (called delays and hidden neurons) to optimize the network performance (weights and bias). The training process can be done with various training (learning) algorithms. The Levenberg-Marquardt (LM) algorithm, the backpropagation (BP) algorithm, the Bayesian regularization (BR) algorithm are examples of most used training algorithms in the literature.

The most essential characteristic of a model for predicting is its ability to generalize. Although generalization competence denotes the model's capacity to perform well on data that was not used to train it, over-fitting prevents model generalization (Schlink *et al*., 2003). To avoid over-fitting, the most frequently used regularization technique is to split the data into three sets (training, testing, and validating).

Different ANN types have been widely used in the literature, but regardless of the type of the utilized ANN, they have some common modeling stages (Figure 33).

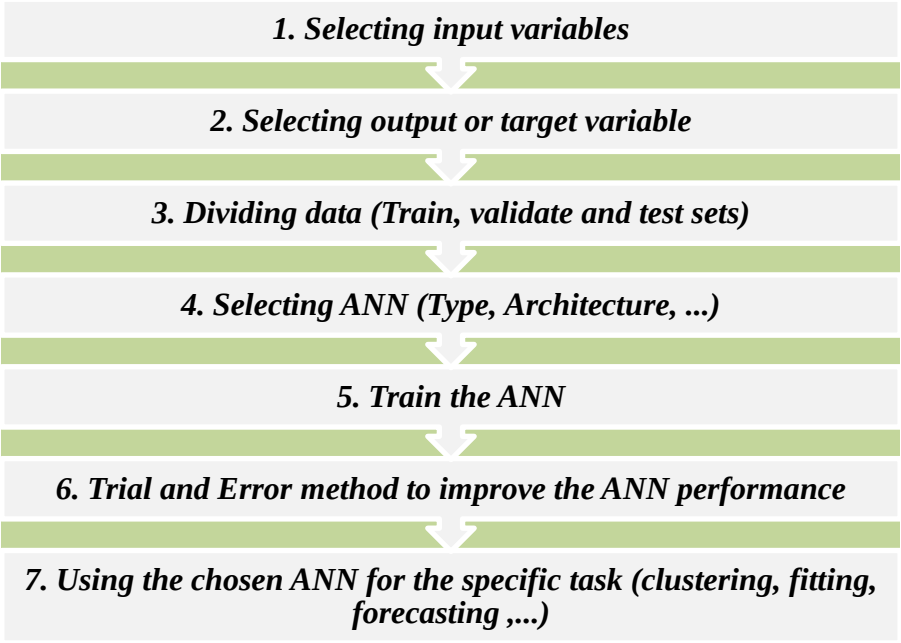



**Figure 33. The common stages of using ANNs.**

Because of its ability to assign meaning to input parameters and to map the inputs to the outputs, the ANN model is an effective modeling technique when relationships between the variables of the underlying physical processes are complex or uncertain (Wei *et al.*, 2001). These neural networks are a nonlinear modeling tool that can manage a large number of inputs to determine one or more outputs (Fogelman *et al.*, 2006).

When compared to conventional models, ANN models can be easily transformed from univariate to multivariate technique, and model complexity may be modified simply by modifying the training process or network design (Rajaee *et al*., 2019). Like the regression models, an empirical evidence or correlation analysis may be used to identify the input variables (Rajaee *et al*., 2019). Furthermore, the results in the literature indicated that ANNs capture the dynamic nonlinear behavior of the time series comparatively better than standard regression models (Rajaee *et al*., 2019).

### 3.2.1. Nonlinear AutoRegressive with eXternal inputs (NARX) neural network

NARX model is a dynamic recurrent neural network (Fig. 34) that encloses several layers (Hayken, 1999). It has previously been applied by many researchers to model nonlinear processes. In 1996, Lin *et al*. stated that NARX network is a powerful modeling and validation tool with a much faster convergence that generalizes much better than other ANNs. There are many applications for the NARX network in representing nonlinear dynamic behaviors.

A NARX model is defined as follows:

$$y(t) = f\big(y(t-1), y(t-2), \dots, y(t-n_y), u(t-1), u(t-2), \dots, u(t-2), \dots, u(t-n_u)\big) \quad (3)$$

Where: *f* is generally unknown and can be approximated; *u (n)* and *y (n)* denote the input and output of the model at discrete time step $n$, respectively.

The first step in a NARX model is to determine the input and output variables. In our study, the output is the Chl-*a* variable, and the input variables are those having the highest permutation importance according to RF model. To ascertain that, different predictor variables combinations were tested. The next step is to set up the network configuration, which consists of determining the number of neurons in the hidden layer and the number of time delays in the input layer to maximize modeling ability. The prediction accuracy (weights and biases) can be improved by adjusting these two parameters (Xu *et al.*, 2019). However, there is no default criterion for determining the optimal structure, we assessed the network's performance with various structures after a training period that was mostly focused on the amount of errors, such as evaluating the MSE and R, as in the RF model. In addition, the error autocorrelation function and input-error cross-correlation function were also checked to evaluate the NARX performance. The autocorrelation error function describes how prediction errors are related in time. For a perfect model of prediction, the difference between the two

errors should be small enough to be statistically insignificant. This would mean that the prediction errors are entirely uncorrelated with each other. This also means that the values of error autocorrelation should mostly be within a certain confidence interval of 95% (Xu *et al.*, 2019).

The input-error cross-correlation function explains how the errors are correlated with the input sequence. For the ideal prediction model, all correlations should be zero, except for the one at zero lag (Markova, 2019).

Three training algorithms, which are the fastest and most adopted in NARX training were tested: Levenberg-Marquardt algorithm, the backpropagation algorithm and the Bayesian regularization algorithm.

Multiple scenarios with different predictor variables combinations (inputs) were tested to simulate the NARX model, and the one with the best performance were used to develop the network for forecasting.

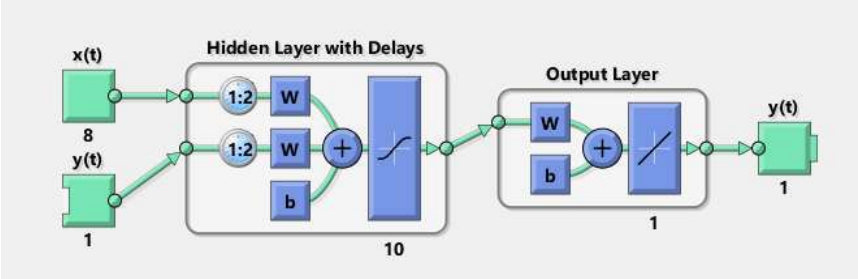



**Figure 34. The structure of the NARX network.**

When using NARX, the network conducts a one-step ahead prediction only after it has been trained. All training is performed in an open loop (also called a series-parallel architecture) including the validation and testing phases. The typical workflow is to fully create the network in an open loop, only when it has been trained (which includes validation and testing steps), it is transformed to closed loop for future predictions (Markova, 2019). NARX performs a one-step-ahead prediction of $y$ ($t$) from historical values of $y$ ($t$) and $x$ ($t$). To do so,

a delay is removed from the network to get the prediction one time step early. The output of the network is then $y(t + 1)$ instead of $y(t)$.

From figure 35, we can see that the network is identical to the previous open-loop network, except that one delay has been removed.

NARX modeling was performed using the MATLAB software MATLAB® software (version 9.3.0.948333 (R2017b), The Mathworks, MA, USA).

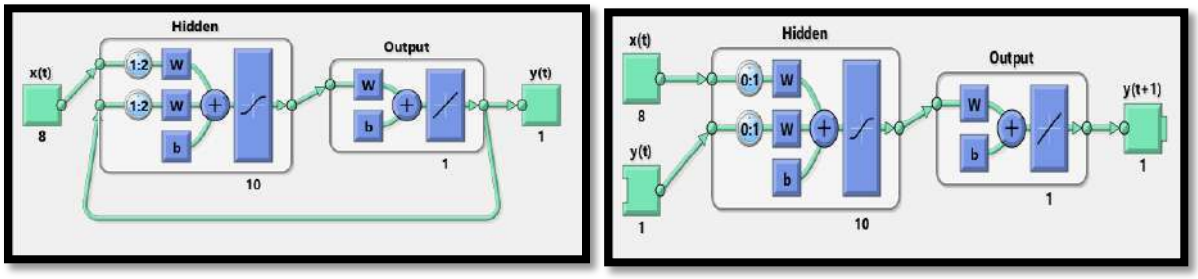



**Figure 35. NARX Neural Network closed loop (left) and one step ahead prediction (right) diagrams.**

# 4. Results and discussion

## 4.1. Nonlinear AutoRegressive with eXternal inputs (NARX) Neural Network

Using nearly three decades of data, the NARX was used to forecast Chl-*a* levels in the North Lagoon of Tunis.

As mentioned above, we used Chl-*a* concentrations as the target, and as inputs, the physico-chemical data, including water temperature, total phosphorus, total nitrogen, Secchi depth, dissolved oxygen, pH and salinity.

The topology with 10 neurons in one hidden layer and 2 lags in the input variables provided the best performance in the prediction of Chl-*a* concentrations among all the tested scenarios.

The Levenberg–Marquardt algorithm as an extensively recognized training algorithm was used (Table 4).

**Table 4. NARX predictions results for Chlorophyll-a between Levenberg-Marquardt, Bayesian Regularization and Scaled Conjugate Gradient algorithms.**

| Algorithm | R | MSE |
|---|---|---|
| **Levenberg-Marquardt** | 0.68 | 0.42 |
| **Bayesian Regularization** | 0.52 | 0.46 |
| **Scaled Conjugate Gradient** | 0.49 | 0.51 |

The network performance was verified by the error autocorrelation function. Figure 36 shows the autocorrelation plot. It indicates that correlations exceeded the 95% confidence limits around zero, meaning the prediction errors are significantly correlated.

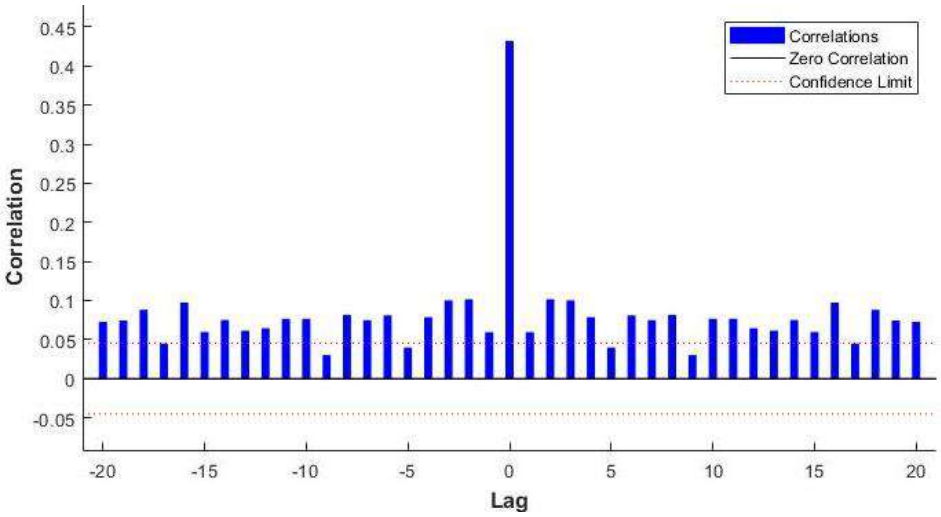



**Figure 36. Autocorrelation plot.**

The training algorithm showed a relative unsatisfying correlation (R=0.68, $R^2$=0.46) between target and output data (Figure 37). The target on the X axis means the observed Chl-*a* concentrations and the output on the Y axis with the equation means the model calculations (model outputs). The correlation R is calculated for the three sets (training, validation and testing), then it is calculated for the whole model.

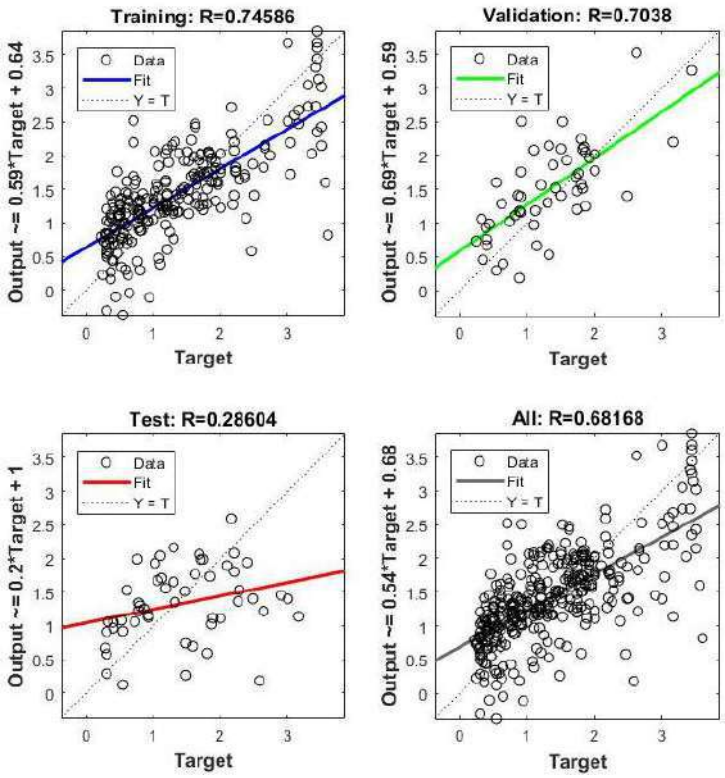



**Figure 37. Correlation between original (target) and predicted (output) Chlorophyll-*a* values obtained with the NARX network.**

The forecasting of the Chl-*a* variation one month ahead indicated about 0.8 µg L-1; it was compared to the observed value (0.5 µg L-1). The forecast is not perfect even though it indicates a relatively normal rate of monthly Chl-*a* variation in the lagoon.

To obtain additional verification on the network performance we verified the error histogram and the input-error cross-correlation function.

Figure 38 exhibits a histogram of errors between the estimated and the actual Chl-*a* data. We can clearly see that the error histogram is biased toward the left and not fitted as a bell-shaped normal distribution, meaning that the errors of the developed model are not normally distributed.

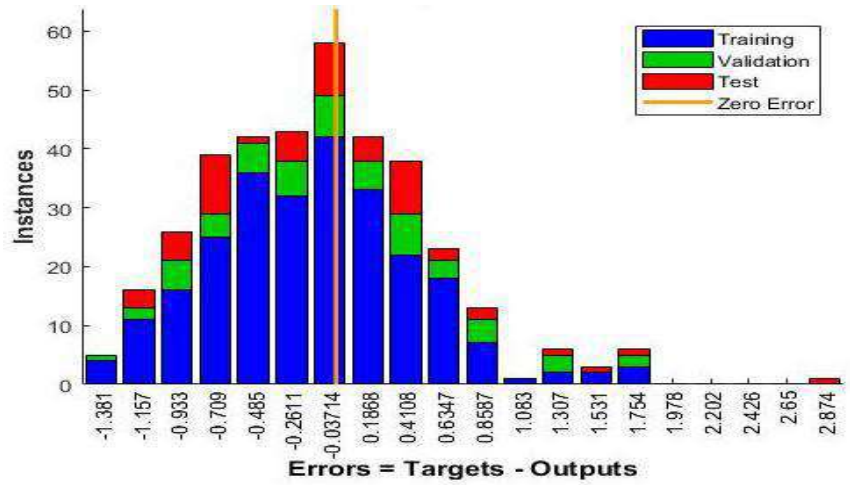



**Figure 38. Error Histogram.**

According to the input-error cross-correlation plot (Figure 39), we see that all the correlations between the errors and inputs exceed the confidence bounds around zero, which means that inputs and errors are correlated.

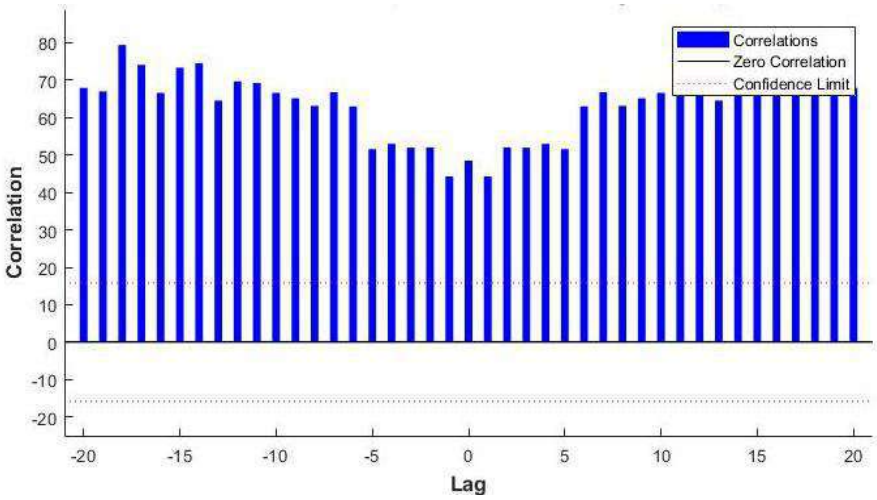



**Figure 39. The input-error cross-correlation plot.**

The network seems to have some performance problems, presenting a relative weak R, and strong correlations while the diagnosis of the errors. Thus, it should be possible to improve the network predictions. To do so, we chose to limitate the number of the variable predictors as external inputs of the NARX. For determining the most important variable predictors, several scenarios have been tested based on Random Forest model from machine learning techniques, and multivariate linear regression model.

## 4.2. Random Forest (RF)

Chl-*a* is one of the most relevant markers of water bodies presence and degree of eutrophication (Lu *et al*., 2016). In the North Lagoon of Tunis, Chl-*a* concentrations range from a minimum of 0.22 µg L-1 and maximum of 3.65 µg L-1. Chl-*a* levels can have complex relationships with both nutrient components (total phosphorus, total nitrogen) and water quality variables (salinity, pH, temperature, dissolved oxygen and Secchi depth) in coastal ecosystems (Jimeno-Sáez *et al.*, 2020). The RF is a suitable technique from ML algorithms when it comes to dealing with complex relations between variables. The RF model was trained on the North Lagoon of Tunis data-352 samples of 7 predictor variables (Secchi depth, dissolved oxygen, temperature, salinity, total nitrogen, total phosphorus and pH) and one target variable, the Chl-*a*. The implementation gave an $R^2$ measure of about 0.62 and MSE equal to 0.28. Figure 40 shows the ranking of predictor variables according to their importance by OOB technique by permutation. Only a few descriptors contributed noticeably to the estimation of the Chl-*a* content namely, Secchi depth followed by the dissolved oxygen and pH.

In aerobic photosynthetic species, Chl-a is the most significant pigment. The depth will impact the strength of sunlight in water and thus the process of photosynthesis (Frolov *et al*., 2012), explaining the strong correlation between the depth and Chl-*a*.

Algae produce oxygen during the day and absorb it during the night. Oxygen consumption also occurs during the process of algae death and decay (Béjaoui *et al*., 2016). In agreement with that, our findings have shown that dissolved oxygen is also associated with Chl-*a* concentrations.

In addition, several studies have demonstrated the strong correlation between Chl-*a* and pH (Menendez *et al*., 2001; Zang *et al*., 2010; Wallace *et al*., 2016).

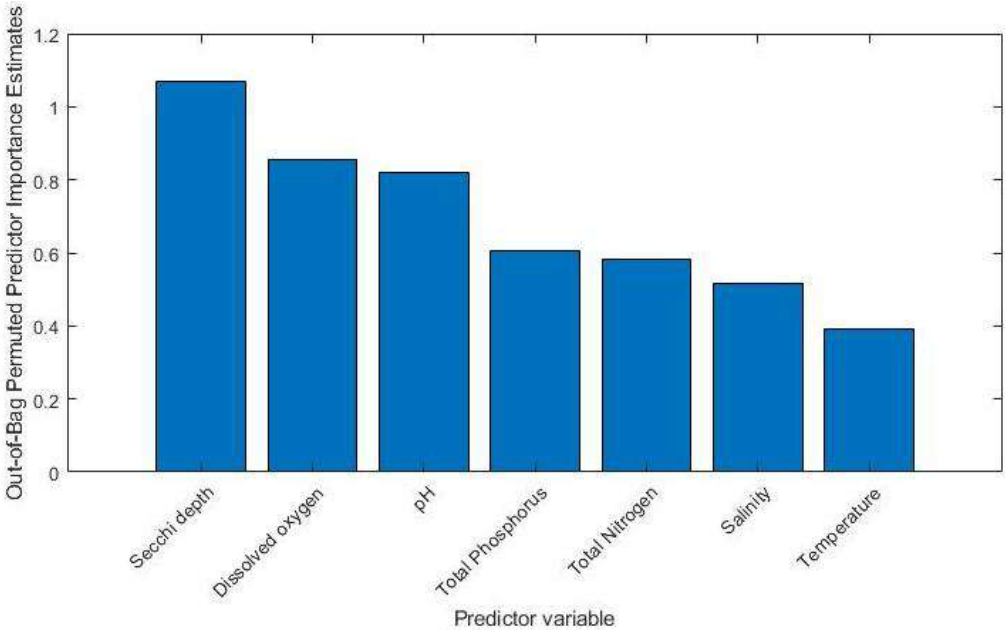



**Figure 40. Predictors importance ranking for the "first" Random Forest model to predict Chlorophyll-*a* content in the North Lagoon of Tunis. The importance of each predictor is measured using the out-of-bag (OOB) technique by permutation due to each predictor.**

In decreasing order of importance, the other predictor variables included in the RF model were: total phosphorus, total nitrogen, salinity and temperature.

A direct comparison (scatter plot) of the observed and predicted Chl-*a* concentrations is shown in Figure 41. The fitted RF model was much better than the one reported by Béjaoui *et al*. (2016) for Bizerte lagoon ($R^2$=0.51), and similar to the one reported by Béjaoui *et al*. (2018) for Ghar el Melh lagoon ($R^2$=0.64). Both lagoons are located in the Mediterranean coast of north Tunisia. Hence, for the North Lagoon of Tunis, the observed Chl-*a* concentrations were more accurately predicted than those of Bizerte lagoon. It is known that for predictive modeling, the number of the observed data is very important for the accuracy of the model (Béjaoui *et al*. 2016). We used long-term of monthly observations that lasted approximately three decades in the North Lagoon of Tunis, which makes the results of the RF accurate in the studied ecosystem.

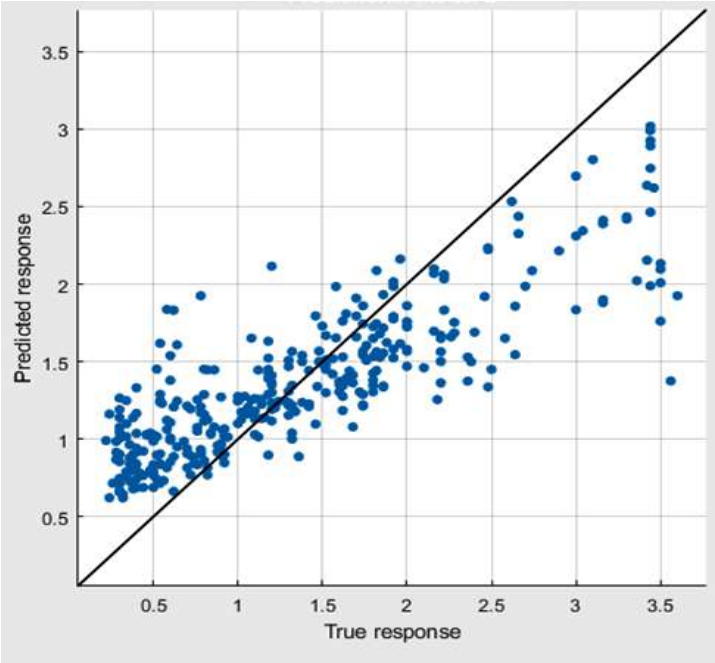



**Figure 41. Random Forest Prediction of the Chl-*a* concentrations using the physico-chemical predictor variables data in the North lagoon of Tunis. Predicted response is Predicted Chl-*a* values and True response is Observed Chl-*a* values.**

For comparison, a Multivariate Linear Regression (MVLR) model was fitted in addition to the RF model. The linear model parameters (Estimate) between predictor variables and Chl-*a* concentrations were almost consistent with the relationships found with the above RF model (Table 3). Both Secchi depth and dissolved oxygen were the two most important predictors explaining Chl-*a* concentration levels, with a p-value $\leq 0.05$. Thus, the linear model quantitatively confirmed the outcomes of the RF model.

P-values and coefficients (estimate) in regression analysis are gathered to tell which relationships in the model are statistically significant. The coefficients describe the mathematical relationship between each independent variable and the dependent variable. The p-values for the coefficients indicate whether these relationships are statistically significant.

SE is the standard error of the coefficients. The t-stat for each coefficient is a value to test the null hypothesis that the corresponding coefficient is zero against the alternative that it is different from zero, given the other predictors in the model. It is to be mentioned that t-stat = Estimate/ SE. The t-stat is then converted to a p-value. A p-value is the probability that the

null hypothesis that all predictors are the same is true. In other words, a lower p-value reflects a value that is more significantly different across predictors. Predictors with significant correlation with Chl-*a* concentrations have p-values $\leq 0.05$.

MVLR gave an $R^2$ of about 0.29. It is obvious that the RF model captures more efficiently the dependency of Chl-*a* concentrations on other variables than the MVLR. The quality of the results is ensured by using the OOB procedure by permutation. We can thus confirm that the RF model could be used to better understand more complex dependencies among variables since it has several advantages over traditional correlative analyses (i.e., a decrease in outlier sensitivity, no implicit assumptions on data distribution).

According to the MVLR, the Chl-*a* concentrations had a significant correlation with water quality variables, as Secchi depth, dissolved oxygen followed by total phosphorus in the study area with a p-value $\leq 0.05$. However, the weakness of its performance suggests that the use of traditional regression methods in the modeling of such a complex process is meaningless, so there is a great need to use more effective techniques (Mjalli *et al.*; 2006).

This fact may support the conclusion drawn by Maier *et al.* (2010) that using a linear approach to define which of the potential input variables have a significant relationship with the model output is not sufficient for the development of ANN models.

All variables were transformed, to normalize their distribution prior to any modeling analyses. However, the transformations did not improve the performance of the MVLR.

The relationship between all variables is strictly non-linear, which is expected and shown in chapter I. Natural ecosystems are governed by several complex processes due to the impacts of hydro-climatic variables such as evaporation, temperature, precipitation, etc. and anthropogenic contribution (Schramm, 1999; Viaroli *et al.*, 2008).

In recent studies of similar lagoons located in the north of Tunisia, Chl-*a* concentrations were found linearly not related to the physico-chemical parameters (Béjaoui *et al.*; 2016; 2018).

We performed all our modeling directly on the original data using ML techniques, known for their abilities to deal with non-linear complex time series processes. Fitting a model directly without transformation is advantageous for forecasting because forecasts are returned on the original scale.

**Table 5. Regression coefficients between Chl-*a* and physico-chemical variables in the North Lagoon of Tunis using linear model (coefficients marked with (*) are statistically significant at p-value < 0.05).**

| | ESTIMATE | SE | T-STAT | P-VALUE |
|---|---|---|---|---|
| **INTERCEPT** | 3.434 | 2.3318 | 1.4727 | 0.14175 |
| ***DISSOLVED OXYGEN*** | -0.12818 | 0.05122 | -2.5026 | 0.012792* |
| ***TOTAL PHOSPHORUS*** | 0.012391 | 0.00753737 | 2.3059 | 0.021713* |
| ***SECCHI DEPTH*** | -0.47539 | 0.17532 | -2.7116 | 0.007032* |
| ***TEMPERATURE*** | 0.0077276 | 0.013261 | 0.58275 | 0.56045 |
| ***PH*** | -0.19153 | 0.21548 | -0.8882 | 0.37472 |
| ***SALINITY*** | 0.027982 | 0.032362 | 0.86464 | 0.38784 |
| ***TOTAL NITROGEN*** | -0.00012694 | 0.0001341 | -0.94663 | 0.34449 |

A second RF model was also fitted by adding two new categorical variables: Station (for an observable spatial dependence) and season (for an observable seasonal dependence). The performance of the RF slightly decreases and attains an $R^2$ of about 0.59.

The addition of a new predictor variable containing spatial information (Station) appears to have little importance on the model simulating Chl-*a* concentrations (Fig. 42). The existence of a strong correlation between Chl-*a* concentrations and Secchi depth showed that depth data itself might be sufficient for demonstrating the spatial variations of Chl-*a* concentrations in the lagoon without including the categorical variable (Station). Moreover, for the RF model, the categorical variable (season) did not really have an important effect on Chl-*a* concentrations. This finding was expected since the variable (temperature) can interfere with the variable (season). Additionally, we may state that this can be due to the climate of Tunis, which stays relatively warm all along the year, including winter.

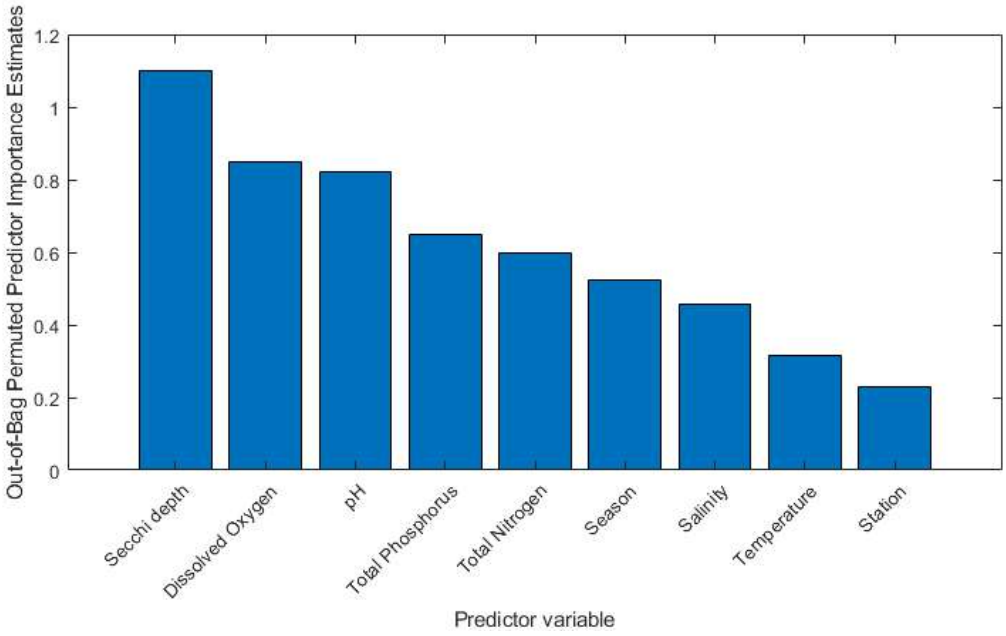



**Figure 42. Predictors importance ranking for the "second" Random Forest model to predict Chlorophyll-*a* content in the North Lagoon of Tunis. The importance of each predictor is measured using the out-of-bag (OOB) technique by permutation due to each predictor.**

RF is a good predictive technique to study the correlations between physico-chemical and/or biological variables in coastal ecosystems. Béjaoui *et al*. (2016) showed that mainly dissolved inorganic nitrogen ($NO_3$) along with dissolved oxygen are the greatest contributors to Chl-*a* content in Bizerte lagoon. Furthermore, in 2018, Bejaoui *et al*. reported that temperature and silicates are the two most strongly correlated variables to the plankton dynamics in Ghar Melh lagoon.

Although, the influence of the predictor variables on Chl-*a* were different in several research works, the dissolved oxygen and Secchi depth generally were among the main variables.

For exemple, Palani *et al*. (2008) applied the ANN model with location variables; orthophosphates ($PO_4$) dissolved oxygen and temperature as the explanatory variables to predict Chl-*a* concentration. Li *et al*. (2017) selected the concentration of total phosphorus and total nitrogen, temperature, Secchi depth, and dissolved oxygen among the most

influential input variables for Chl-*a*. Furthermore, Kuo *et al*. (2007) defined the Chl-*a* model by the input of month, temperature, pH, Secchi depth, suspended solids (SS), $PO_4$ and $NO_3$.

It is important to highlight that the difference in the predictor variables selection between the previous ecosystems with the North Lagoon of Tunis can be explained by the ecosystem specificities, as dimensions of water masses, different eutrophic states, water depth and communication with the sea. In addition, various modeling approaches, in addition to different field works and laboratory analysis techniques may have contributed to these differences.

## 4.3. Nonlinear AutoRegressive with external inputs (NARX) Neural Network

In this study, four scenarios with different input combinations of the predictor variables are tested for estimating and forecasting Chl-*a* concentration values in the North Lagoon of Tunis using the NARX network.

The first input scenario (S1) considered all parameters we had as inputs without selection. The second scenario (S2) included only the three most important predictor variables according to the RF model. The third input scenario (S3) included only the most highly correlated parameters according to the MVLR. The last scenario (S4) included only the two most important predictor variables according to the RF model. Summarized results of predictor variables selection are shown in Table 4.

**Table 6. Summarized results of predictor variables selection.**

| Algorithm | N. of predictor variables (inputs) selected | Predictor variables selected | Input scenario |
|---|---|---|---|
| **NARX without inputs selection** | 7 | All | S1 |
| **NARX with three most important predictor variables selected with RF** | 3 | Depth<br>Dissolved oxygen<br>pH | S2 |
| **NARX with the correlated predictor variables selected with MVLR** | 3 | Depth<br>Dissolved oxygen<br>Total phosphorus | S3 |
| **NARX with two most important predictor variables selected with RF** | 2 | Depth<br>Dissolved oxygen | S4 |

The NARX models with the four input scenarios described in Table 4 were developed to simulate the Chl-*a* concentrations. The four versions of each model represent four substantially different chlorophyll-*a* models, due to the different combinations of variables used as predictors. Different ML models are compared based on statistical indices such as R, $R^2$, MSE, etc (Jimeno-Saez., 2020)**.** These performance measurements are summarized in Table 5 for the NARX network. Different parameters are tried for each NARX model and the best one; with the minimum MSE and maximum R and $R^2$ in selected for the forecast of Chl-*a* task.

The topology with 10 neurons in one hidden layer and 2 lags in the input variables provided the best performance in the prediction of Chl-*a* concentrations among all the scenarios. Our study considered that a proportion of 70% training, 15% validation and 15% test is a favorable implementation.

The Levenberg–Marquardt algorithm, an extensively recognized training algorithm, was used for minimizing nonlinear functions. Training automatically stops when generalization stops

improving, as indicated by an increase in the mean square error of the validation samples (Mammadli, 2017; Xu *et al*., 2019).

**Table 7. Performance of Chl-*a* estimation from NARX models based on four different input scenarios.**

| Scenarios | R | $R^2$ | MSE |
|---|---|---|---|
| **S1** | 0.68 | 0.46 | 0.42 |
| **S2** | 0.75 | 0.56 | 0.31 |
| **S3** | 0.74 | 0.55 | 0.40 |
| **S4** | 0.79 | 0.62 | 0.31 |

The comparative results between the four versions of the NARX model reveal that the NARX with two inputs selected by the RF algorithm yielded the best accuracy among all the developed NARX models in term of higher R and $R^2$ and lower MSE values (R= 0.79; $R^2$=0.62; MSE=0.31).

With three inputs, selected also according to the RF model, the S2 scenario is the second most accurate model with a performance close to the best one.

Because simulating time and over-fitting risks increase with the number of predictor variables in predictive modeling, a good practice is to create a model using as few predictor variables as possible (Jimeno-Saez at al., 2020).

We now present one step-ahead (a month) forecasting results for the eutrophication indicator considering the three datasets: Chl-*a* concentrations as the target, using Secchi depth and dissolved oxygen as external inputs.

Most of the NARX model errors were very close to zero and fall within the confidence interval (Fig.43), therefore the autocorrelation errors were negligible.

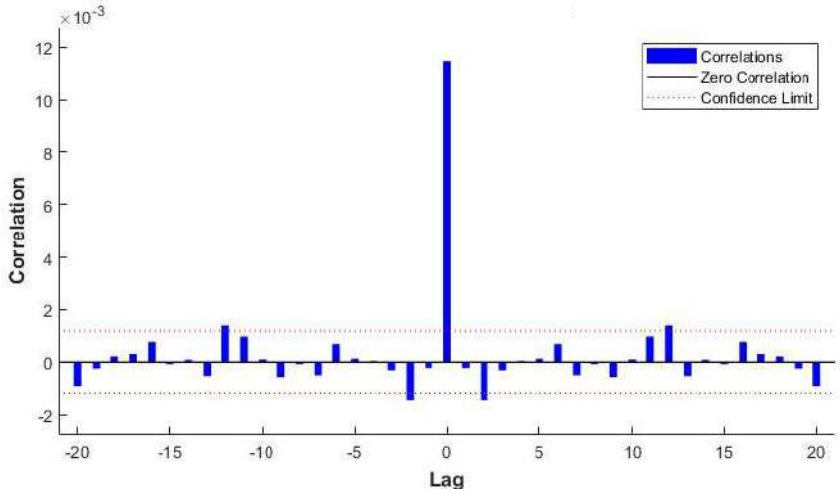



**Figure 43. Autocorrelation plot.**

In general, the input-error cross-correlation plot (Figure 44) showed that all the correlations fell within the confidence bound. Furthermore, the obtained results (Fig.45) show an overall correlation of R= 0.79 between the actual data (targets) and the predicted values (the outputs). The error histogram was checked. It presents a closely bell-shaped normal distribution of the errors (Fig.46). Given that, we can conclude that, residuals of the NARX model are uncorrelated and normally distributed.

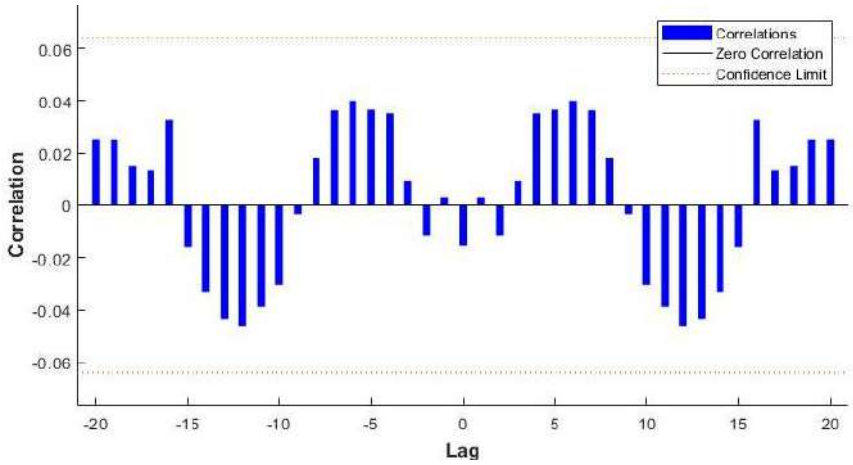



**Figure 44. Cross-correlation plot.**

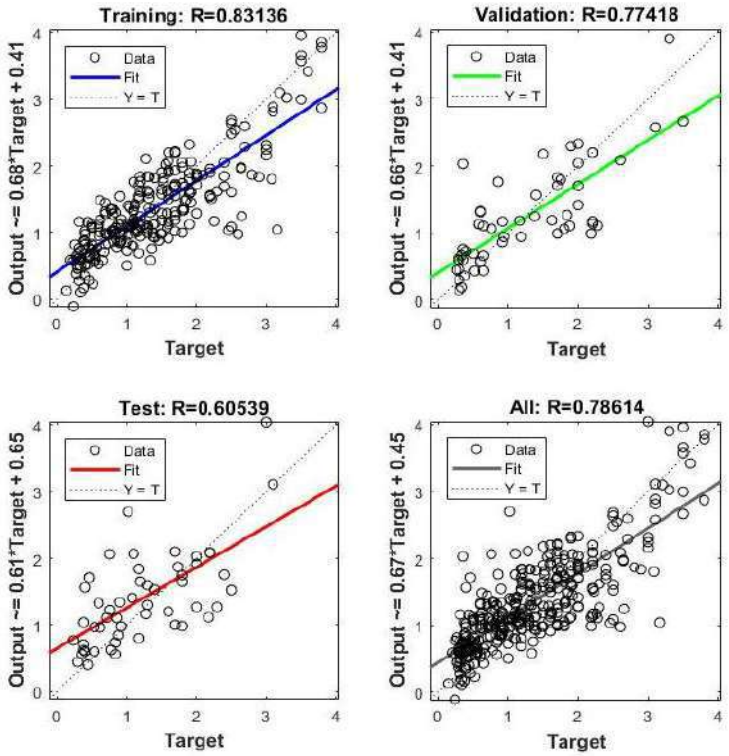



**Figure 45. Correlation between original (target) and predicted (output) Chlorophyll-*a* values obtained with the NARX network.**

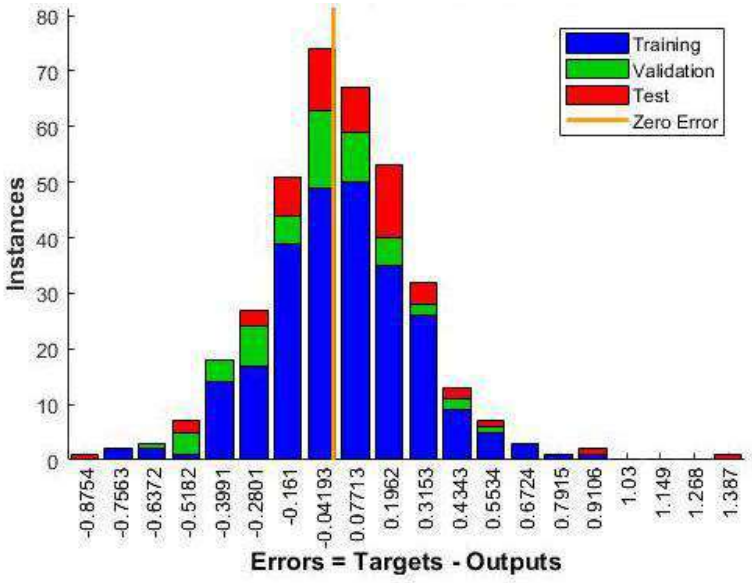



**Figure 46. Error Histogram of the NARX network.**

The model's fitness is described in figure 47 and a visual comparison of Chl-*a* concentrations predictions with respect to the observed data is shown. There is a fairly good match between

the observed values and the fitted values. The NARX network was able to predict the high variability of Chl-*a* concentrations, therefore, the fitted model seems to be mathematically accurate and the NARX could be used on a new data set. Given its effectiveness, multiple studies used neural networks techniques to model the Chl-*a* contents as an eutrophication indicator in coastal ecosystems. We can mention the study of Nazeer *et al*. (2017), who suggested, using ML methods, such as ANN for a more accurate and efficient routine monitoring of coastal water quality parameters, particularly Chl-*a*, in a coastal area of Hong Kong. In another study carried out in the Mar Menor lagoon in Spain, a Multilayer Neural Networks have been used for the eutrophication modeling, considering Chl-*a* as the eutrophication indicator (Jimeno-Sáez *et al.*, 2020). In 2003, Lee *et al*. (2003) used back-propagation learning algorithm for training the ANN to predict the algal bloom dynamics of the coastal waters of Hong Kong using a 4-year set of phytoplankton abundance data. Lu *et al.* (2016) used a back-propagation ANN model for the prediction of Chl-*a* concentrations in Lake Champlain in China.

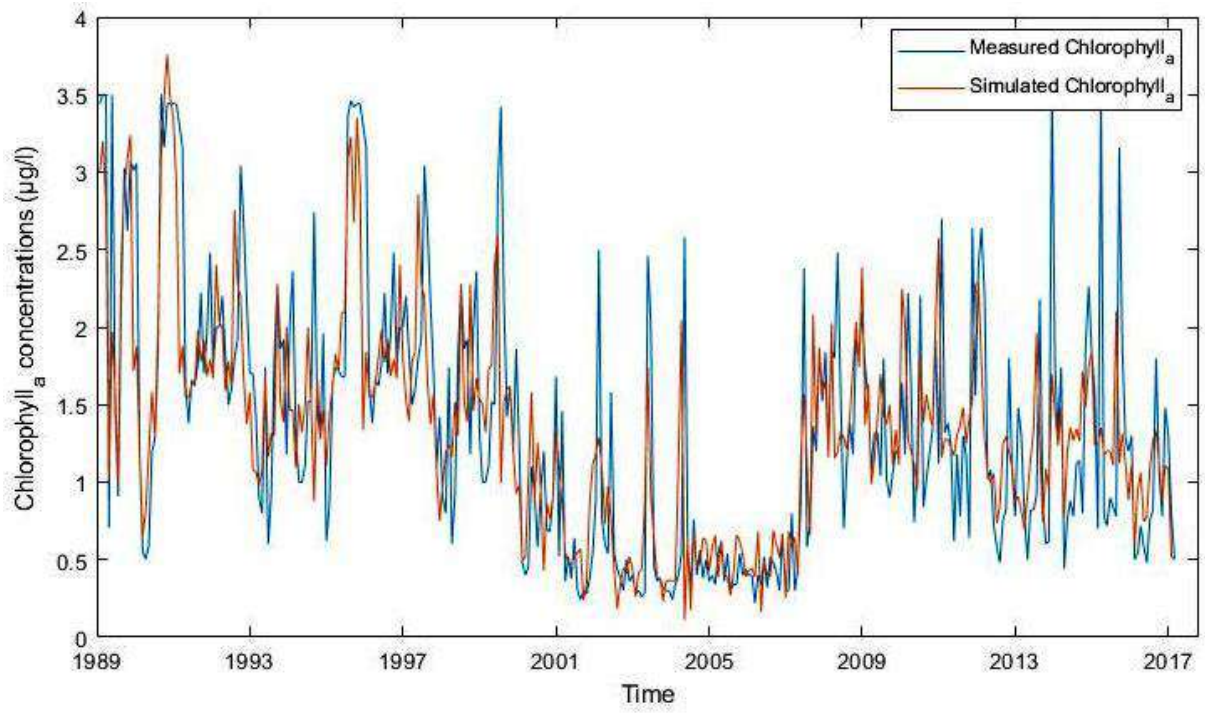



**Figure 47. Observed and modeled Chlorophyll-*a* concentrations using NARX Network.**

By comparing the NARX results (R= 0.79; $R^2$= 0.62) with the MVLR including just the two most important predictors for the Chl-*a* concentrations, the performance of the latter linear model decreases further and attains an $R^2$ of about only 0.2. We can thus confirm that the relationship between the variable predictors and Chl-*a* concentrations is obviously non-linear and the use of the NARX, RF and generally ML techniques, is adequate for forecasting Chl-*a* contents in the studied lagoon.

The forecasting of the Chl-*a* content one month ahead gave a value about 0.51 µg L-1, which was close to the observed value (0.5 µg L-1). These values are practically similar, which indicates a relative normal Chl-*a* level of monthly variation in the lagoon and the accuracy of the developed NARX model.

## 5. Station 5

Since Station 5 was found the area the most affected by the eutrophication, due to its position (the furthest from the sea), it is important to enhance our knowledge of the eutrophication process and the connections between the lagoon's water quality indicators in this sensitive area.

In this section, I present the results from training using only historical data of Station 5, with the same machine learning approaches (Random Forest and Nonlinear AutoRegressive with eXternal inputs neural network), as in this chapter.

Figure 48 illustrates the ranking of predictor variables according to their importance by OOB technique by permutation. It shows clearly that total phosphorus along with salinity followed by the total nitrogen are the variables that best predict Chl-*a* concentrations at Station 5.

This result is expected, since the greater concentration of Chl-*a* at Station 5 compared to the other Stations corresponds to the high level of all nutrients detected in this section of the lagoon. These waters originate from a location (Station 4) that was unaffected by the dredging

work carried out as part of the lagoon restoration project. This area of the lagoon is distinguished by a nutrient-rich bottom and the permanent renewal of the lagoon waters is allowing the gradual washing of the nutrient reserves from the sediment.

Total phosphorus and total nitrogen are nutrients that every aerobic photosynthetic organism requires for growth. Several studies in different coastal ecosystems have demonstrated the significant link between Chl-*a* and total phosphorus and total nitrogen concentrations (Souchu *et al*., 2010; Bennett *et al*., 2017; Ruzafa *et al*., 2019).

Salinity is one of the most significant factors influencing the dispersal of living creatures, among others photosynthetic organisms. Multiples researchers demonstrated the strong correlation between the Chl-*a* and the salinity in aquatic ecosystems (Håkanson and M Eklund 2010; Desmit *et al*., 2015).

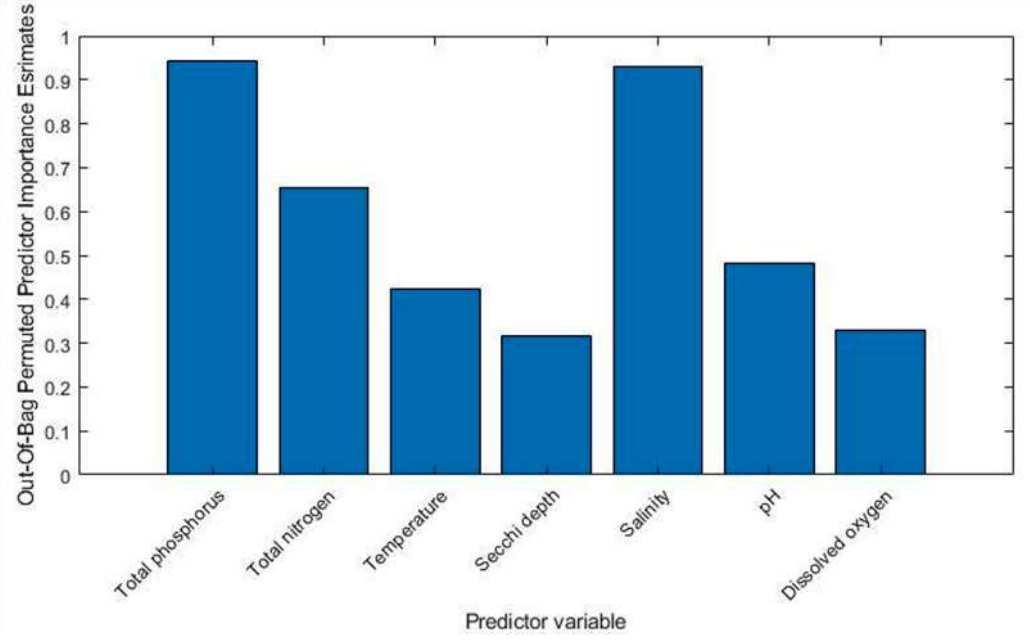



**Figure 48. Predictors importance ranking with Random Forest model in Station 5. The importance of each predictor is measured using the out-of-bag (OOB) technique by permutation due to each predictor.**

Here the Random Forest (RF) model produced an $R^2$ of around 0.55 (table 6). This result is close to the result achieved by modeling RF with monthly dynamics of each variable across

the whole lagoon ($R^2$= 0.62). This slight decrease in the model performance maybe attributed to the data that is qualified to be more heterogeneous and have more enhanced nonlinear relationships in Station 5.

Despite this small decrease in the prediction accuracy, RF is identified as an effective technique to select the specific variables that are the most related to Chl-*a* concentrations at the Station 5 of the North Lagoon of Tunis.

In order to forecast the Chl-*a* concentrations, one step ahead early in Station 5, a nonlinear autoregressive with external inputs neural network (NARX) is performed.

The implemented NARX considers the variables that best predict the Chl-*a* concentrations (total phosphorus and salinity) as inputs and the Chl-*a* as the target.

Among all scenarios, the architecture with 10 neurons in one hidden layer and two delays in the input variables offered the highest performance in predicting Chl-*a* concentrations.

Our study considered that a proportion of 70% training, 15% validation and 15% test is a favorable implementation.

Since the Levenberg–Marquardt algorithm is famous for reducing nonlinear functions (Mammadli, 2017; Xu *et al*. 2019), it is not surprising to determine that it is the best method to use here.

The variations between the observed and simulated data from Station 5 are clearly less important than the fluctuations recorded between the observed data in Station 5 and the simulated data of the whole lagoon performed previously in this chapter (Figure 49). It addition, figure 50 represents the models fitness and the comparison of Chl-*a* concentration predictions in the lagoon and at Station 5 versus the observed data at Station 5. The latter and the fitted values at Station 5 are reasonably close and predicts more effectively than the

NARX performed for the whole lagoon. Considering that, it is of a big importance to develop the NARX for Station 5 and look closely at the Chl-*a* levels in this area of the lagoon.

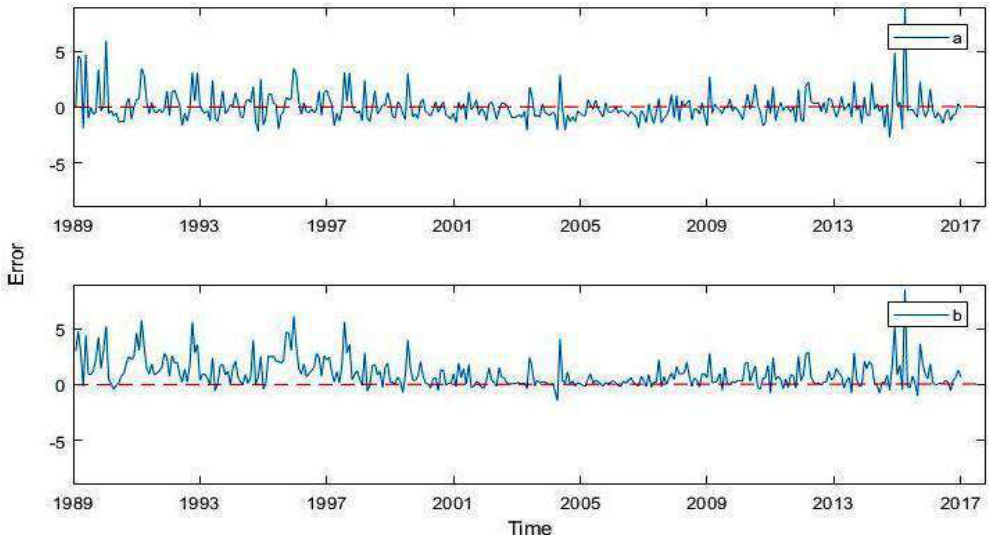



**Figure 49. (a) Variations between observed Station 5 Chl-*a* data and simulated Station 5 Chl-*a* data (mean = 0.016 +/- 0.003), and (b) variations between observed Station 5 Chl-*a* data and simulated Chl-*a* data from the lagoon data (mean = 0.99 +/- 3.13).**

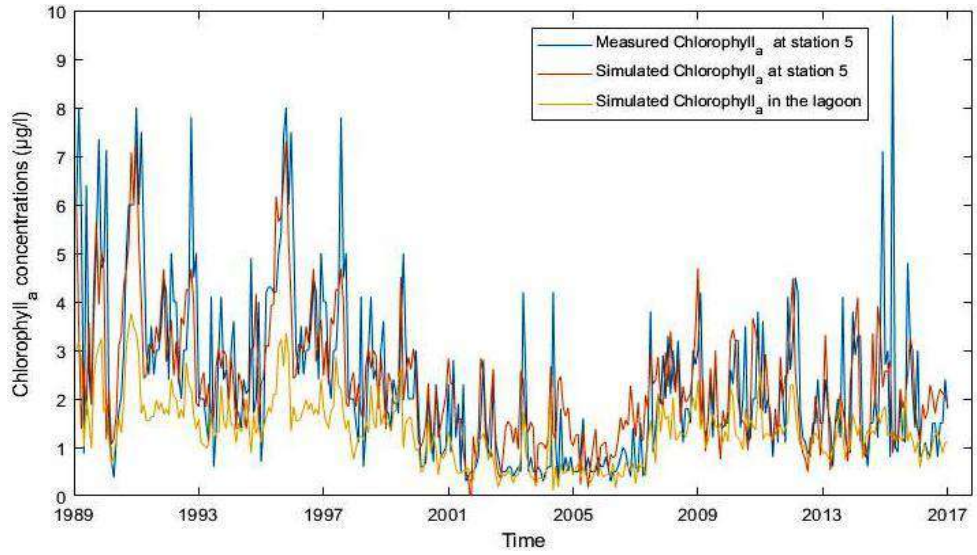



**Figure 50. Measured and simulated Chl-*a* concentrations using NARX network at Station 5 and in the North Lagoon of Tunis.**

The NARX network was able to predict Chl-*a* concentrations with satisfactory results and the fitted model appears to be mathematically accurate. It can be observed that the NARX has slightly underestimated at some high concentrations of Chl-*a*, while adequately estimated Chl-*a* at reasonable concentrations. This may be due to the training of the network with

mostly reasonable concentration data. In addition, the data sets' underestimate is most likely due to the non-homogeneous character of the input and output water quality variables. These data were collected over a 30-years period, resulting in a considerable variance in the values of water quality indicators in this study.

The error autocorrelation function is used to assess the network performance (Figure 51). The autocorrelation plot indicates that almost all correlations lie within the 95% confidence limits around zero, which indicates that the prediction errors are significantly uncorrelated.

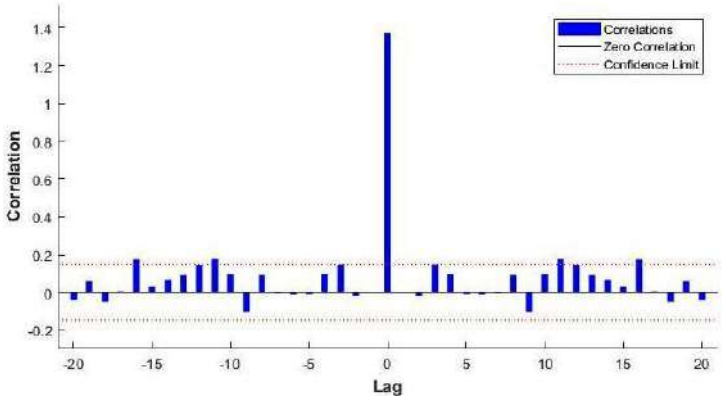



**Figure 51. Autocorrelation plot.**

**Table 8. Comparison between the lagoon results and the Station 5 results.**

| Techniques | Lagoon | Station 5 |
|---|---|---|
| **Random Forest** | $R^2$= 0.62<br>Predictor variables:<br>Depth/ Dissolved oxygen | $R^2$= 0.55<br>Predictor variables:<br>Total Phosphorus/ Salinity |
| **NARX** | R= 0.79<br>MSE= 0.31 | R= 0.73<br>MSE= 0.40 |

Overall, a high correlation (R= 0.73) is observed between in situ measured and estimated Chl-*a* concentrations in Station 5 (Figure 52). This result is slightly lower than the previous developed NARX (R=0.79) and the non-Stationarity and heterogeneity of the time series can explain this small decrease (table 6), especially at Station 5, where there are high fluctuations.

Nevertheless, the predictions of the Chl-*a* content one month ahead gave a value about 1.9 µg L-1, which was close to the observed value (1.7 µg L-1). These results are quite similar, indicating a typical monthly variation in Chl-*a* levels in the lagoon at Station 5, which shows the good performance of the proposed NARX model.

For comparison, we simulated a NARX network with total phosphorus, salinity and total nitrogen as inputs. The performance decreases further and attains R= 0.69.

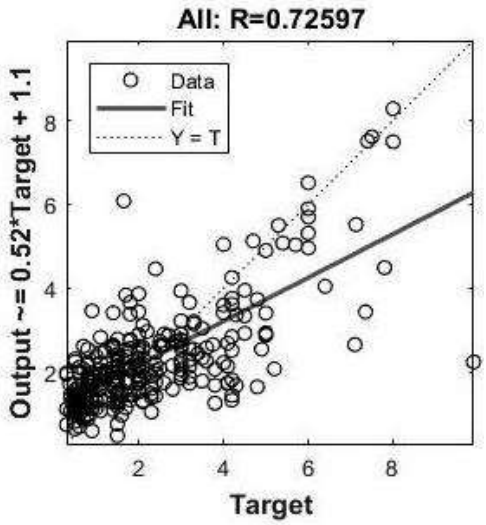



**Figure 52. Correlation between original (target) and predicted (output) Chl-*a* values obtained with the NARX network in Station 5.**

To quantify the spatial heterogeneity in the lagoon, a comparison is made between model results simulated using monthly dynamics of each variable in the whole lagoon and model results simulated exclusively with data from Station 5. It is demonstrated that using only a few predictor variables, the NARX created was able to forecast Chl-*a* concentrations dynamics quite effectively in Station 5.

It is important to mention, that the North Lagoon of Tunis offers great possibilities for fishing, because the high salinity of its waters has made plankton flourish feeding mollusks and fish in large quantities. There are fisheries at the exit of the lagoon to catch the fish as they leave the

ecosystem and head for the open sea. So, it is of a great importance to control the water quality in this area of the lagoon.

The main factors to the eutrophication evaluation and forecasts in Station 5 are total phosphorus and salinity. According to our results, it is recommended to regularly monitor total phosphorus and salinity in the south part of the lagoon, especially at Station 5. In the north part of the lagoon (Stations 1, 2 and 3), it is suggested to control the Secchi depth (transparency) and the dissolved oxygen. This variable selection is a crucial step in order to reduce the cost and time consuming of field monitoring and laboratory analysis.

## 6. Conclusion

The approach proposed in the current chapter relies on a combination of ML methods, using NARX neural network and RF model to predict and forecast Chl-*a* concentrations in the North Lagoon of Tunis. Mainly Secchi depth along with dissolved oxygen are the greatest contributors to this eutrophication assessment and forecasting. Our results agree well with findings from other studies carried out on Mediterranean coastal lagoons. It's worth mentioning that Secchi depth and dissolved oxygen are very practical variables to measure, without the need of extra laboratory analysis.

The NARX developed was able to predict Chl-*a* concentration dynamics fairly well using minimal input predictor variables (R= 0.79). Our results show that complex behavior in the eutrophication process could be modeled using the NARX technique and, some extreme values were successfully estimated.

The same approaches were simulated on historical data of the Station 5, which is the most affected area with eutrophication in the lagoon. Total phosphorus and salinity were demonstrated to be the variables that best predict the Chl-*a* levels at Station 5, and the created NARX was able to forecast the Chl-*a* concentrations quite effectively (R= 0.73).

The findings verified the relevance and usefulness of intelligent modeling as a tool that is simple, quick, easy to use, and inexpensive. The developed model can be used to (1) estimate Chl-*a* concentrations when the real value is unavailable (2) simulate alternative water quality scenarios over extreme ranges of input and output parameters.

It is important to mention that despite the important amount of the observed data (approximately three decades) used for developing the NARX, it has a very short computational time.

In the wider context of the study of coastal lagoons and other transitional ecosystems, our approach could be used to assess and predict the eutrophication process of these natural environments and help in assisting civil authorities, engineers, economists, investors, and other interested stakeholders in making decisions.

# Chapter III: Long time Chl-*a* forecasting in the North lagoon of Tunis using Box and Jenkins methodology

## 1. Introduction

The focused monitoring of major water pollution indicators may detect eutrophication episodes within coastal settings (Chen *et al*., 2015). Chl-*a* measurements are used to estimate the amount of phytoplankton biomass in the water, the potential algal bloom, and hence the degree of environmental eutrophication (Tian *et al*., 2017).

The mechanisms and processes involved in the initiation and outbreak of eutrophication are important to control. Anticipating possible eutrophication episodes is an important emergency management measure (Villanoy *et al*., 2005; Dippner *et al*., 2011; Recknagel *et al*., 2013, 2014), which, in turn, has stimulated the development of ecological forecasting models.

As mentioned earlier in this manuscript, data driven models based on auto-regression, multivariate regression, piecewise regression or artificial neural networks (ANNs), were rapidly developed and applied to eutrophication forecasting. Despite the lack of an explicit function to define the model system, data-driven methods are effective at forecasting data (Chen *et al*., 2015). Oh *et al*. (2007) applied two ANN models to identify temporal phytoplankton community patterns. Wang *et al*. (2010) developed an ANN model to predict cyanobacterial blooms based on weather conditions for use as an early warning system.

Regression models are set to identify governing factors and to establish their approximate relationship to dependent variables (Cui *et al*., 2007; Onderka, 2007; Davis *et al*., 2009; Wilhelm *et al*., 2011) using data samples to estimate model parameters and evaluate model performance. Ouchi (1982) applied principal component analysis to develop a multivariate prediction model for algal biomass in northern Hiroshima Bay, Japan. Lui *et al*. (2007) developed a vector-based auto-regression model for algal bloom forecasting in Hong Kong

waters. For Stationary systems, these models typically show a high predictive capability. However, eutrophication is event-driven and characterized by non-Stationary characteristics (Hasting, 2001; Onderka, 2007; Paerl and Huisman, 2008; Recknagel *et al*., 2013). More importantly, aquatic ecological data are usually sparse and incomplete, lacking either in hydro-environmental or biological parameters. Such data makes it difficult to create multivariate regression models, despite the availability of strategies for dealing with missing data (Donner, 1982).

Box and Jenkins (1976) techniques have been widely employed in a variety of disciplines. Box and Jenkins is a dynamic, computer-based iterative process that generates an autoregressive, integrated moving average model (ARIMA), optimized for seasonal and trend variables (Gaynor and Kirkpatrick, 1994).

ARIMA model is being used in different studies. For instance, to predicting water levels in Lake Malawi (Makwinja *et al*., 2017), or water salinity in Apalachicola bay in Florida (Sun and Koch, 2001) and to forecasting sulphur dioxide in Tehran (Hassanzadeh *et al*., 2009). Chen *et al*. (2015) have developed an ARIMA model to predict daily Chl-*a* concentrations in Taihu Lake (China) and demonstrated its effectiveness in comparison to a multivariate linear regression (MVLR). The seasonal, autoregressive, integrated moving average (SARIMA) model is composed of ARIMA model including seasonal component of the time series data. SARIMA is very frequently used for monthly time series that exhibit a seasonal pattern (Prista *et al*., 2011). Chl-*a* is a parameter known to be related to temperature (Tizro *et al*., 2014), which have seasonal characteristics. For that reason, a SARIMA model is implemented to handle the characteristics of the Chl-*a* seasonal variations, which improves the prediction accuracy (Tizro *et al*., 2014). Furthermore, the SARIMA model requires just one input variable, indicating greater application.

The objective of this chapter was indeed to investigate the applicability of a SARIMA model in algal bloom forecasting, using Chl-*a* concentrations as a eutrophication indicator using a period of approximately three decades (January 1989 - April 2018) of retrospective Chl-*a* data in the North Lagoon of Tunis.

In comparison, SARIMA was also applied to forecast Chl-*a* concentrations at Station 5 of the lagoon.

## 2. Time Series

A time series is, at its most basic definition, a set of measurements taken progressively through time.

Time series analysis was used in the Kyoto Protocol for lowering greenhouse gas emissions. Scientific research, economics, and time series analysis are all used to justify reducing greenhouse gas emissions. Decisions made in the coming years will have an impact on the planet's future (Cowpertwait and Metcalfe, 2008).

Time series techniques are employed in operational decisions on a daily basis. Gas suppliers in the United Kingdom, for example, would place orders for gas from offshore sources one day ahead of delivery (Cowpertwait and Metcalfe, 2008). Geophysicists are constantly studying the earth's shaking or trembling in order to forecast potentially coming earthquakes (Cowpertwait and Metcalfe, 2008). An electroencephalogram records brain waves produced by an electroencephalograph in order to detect a neurological disease (Cowpertwait and Metcalfe, 2008). In addition, we can mention the recent events when scientists are using the daily number of Coronavirus new cases and deaths to predict how the pandemic will evolve in the near and far future, so they can make the right decisions about this crisis.

Obviously, there are several justifications for documenting and evaluating time series data. Among these are the capacities to gain a better knowledge of the data generation process and anticipate future values for optimal system control (Cowpertwait and Metcalfe, 2008).

## 3. Common approaches to time series

When it comes to evaluating time series data, there are four main techniques.

Briefly, decomposition models, smoothing methods, autoregressive models, and Box and Jenkins methodology are described here for time series analysis and forecasting.

- In preparation for prediction, decomposition techniques give strategies for dividing the time sequence into trends, cycles, seasons, and irregular components [6].

- Smoothing methods are extrapolation techniques based on moving average (MA) [6]. The moving average model is the most fundamental method to time series modeling [6]. The next observation is just the mean of all previous observations. Despite its simplicity, this model may be surprisingly accurate, and it serves as an excellent starting point. Otherwise, the moving average can be utilized to spot intriguing data trends. We may create a window to smooth the time series and emphasize various trends using the moving average methodology [6].

- Briefly, autoregressive models are essentially a regression of the time series onto itself [6].
- The Box and Jenkins techniques combine simpler models to create a sophisticated model for time series with non-Stationary features and seasonality. There is the ARIMA model, which differentiates the series to Stationarity and then combines the moving average (MA) with autoregressive (AR) parameters to yield a comprehensive model amenable to forecasting [4]. There is the SARIMA when dealing with seasonality and finally the ARMA method, when dealing with a Stationary time series.

Further in this chapter we'll discuss more in details the Box and Jenkins methodology.

# 4. Box and Jenkins

## 4.1. Presentation

Box and Jenkins is a dynamic, computer-based iterative method that creates an auto-regressive, integrated moving average model that is applicable for seasonal and trend factors

(Gaynor and Kirkpatrick, 1994). The technique is suitable for medium to long-term time series with at least 50 observations (Wei, 1990).

The ability to deal with complicated dynamic processes, its flexibility in processing dependent time series data, its advanced computational and mathematical methods, its functionality in uncertainty analysis and the simplicity of its execution (Lu et AbouRizk, 2008), made Box and Jenkins method one of the best techniques to create a model as a forecasting tool .

Box and Jenkins methodologies provide the most reliable forecasting models for any data set (Gaynor and Kirkpatrick, 1994). Armstrong's comparison test (Armstrong,1985) on the ranking of extrapolation methods (from the highest rank as 1 to the lowest rank as 5) for both short-range and long-range graded Box and Jenkins methods in terms of cost, comprehensibility, and forecast accuracy as 1.5 for short-range forecast accuracy and 2 for long-range forecast accuracy (Lu et AbouRizk, 2008).

Box and Jenkins methods generally start with the most recent observations and then analyze recent forecasting errors to determine the appropriate modifications for future periods. In doing so, they enable a more flexible imitation of a certain dynamic pattern or seasonality while making timely adjustments to error levels (Lu et AbouRizk, 2008).

Box and Jenkins models are capable of handling dependent time-series data that are considered not suitable for other approaches. A regression model, for instance, has a standard assumption that the term for error should be statistically independent. In fact, often related data are either dependent or correlated with each other (Lu et AbouRizk, 2008).

Box and Jenkins' technique is based on the fact that the process that generated the time series can be approximated using an ARMA model if it is Stationary or an ARIMA model if it is non-Stationary (Lu et AbouRizk, 2008).

## 4.2. Stationarity

A time series is considered to be Stationary if there is no systematic change in mean (no trend). In other words, the characteristics of one segment of data are relatively similar to those of any other segment (Cowpertwait and Metcalfe, 2008).

The time plot should give an idea if the time serie is Stationary or not, but it's important to double check the Stationarity in a mathematical way. To do so, a test is done called the Augmented Dickey Fuller Test (ADF).

In statistics, the ADF test examines the null hypothesis that the time series is non-Stationary. The alternative hypothesis is that the time series is Stationary [7].

- **Null Hypothesis (H0)**: If it is not rejected, it shows that the time series is non-Stationary.
- **Alternate Hypothesis (H1)**: The null hypothesis is rejected; it implies that the time series is Stationary.

The p-value from the test is used to interpret this result. A p-value less than a certain threshold (5%) indicates that we reject the null hypothesis (Stationary); conversely, a p-value greater than the threshold indicates that we fail to reject the null hypothesis (non-Stationary).

- **P-value > 0.05**: Fail to reject the null hypothesis (H0), the data is non-Stationary.
- **P-value <= 0.05**: Reject the null hypothesis (H0), the data is Stationary.

## 4.3. The Autocorrelation Function (ACF)

The concepts of correlation are very important in time series analysis. In particular, we can examine the correlation structure of the original data to help us identify possible form(s) of (none) Stationary models [1]. Autocorrelation is the correlation of a variable with itself at

differing time lags. Autocorrelation between any two observations only relies on the time lag (*h*) between them [1]. For Stationary processes define:

$$Cov\ (y_t, y_{t-h}) = \gamma_h$$

Lag-*h* autocorrelation is given by:

$$\rho_h = Corr(y_t, y_{t-h}) = \frac{\gamma_h}{\gamma_0}$$

This function plays an essential role in the analysis of data aimed at determining the degree of the lag in a moving average (MA) model [1].

The usage of this function was presented as part of the Box and Jenkins approach to time series modeling, in which displaying the autocorrelation function allowed one to identify the proper lags **q** in an MA (**q**) model or an extended ARIMA (**p,d,q**) model (Omer, 2010).

## 4.4. Partial autocorrelation function (PACF)

The partial autocorrelation function (PACF) measures the linear correlation of a series $x_t$ and a lagged version of itself $x_{t+k}$ with the linear dependence of $\{x_{t-1},x_{t-2},\ldots,x_{t-(k-1)}\}\{x_{t-1},x_{t-2},\ldots,x_{t-(k-1)}\}$ removed [1].

This function plays an important role in data analysis aimed at identifying the extent of the lag in an autoregressive (AR) model [1]. The use of this function was introduced as part of the Box and Jenkins approach to time series modeling, whereby plotting the partial autocorrelation function one could determine the appropriate lags **p** in an AR (**p**) model or in an extended ARIMA (**p**,**d**,**q**) model (Omer, 2010).

In more plain words, the Autocorrelation Function and the Partial Autocorrelation Function are functions that express information useful in determining the orders p and q of an ARIMA (p,d,q) model (table 7).

**Table 9. ACF and PACF in identifying *p* and *q* [1].**

| Conditional Mean Model | ACF | PACF |
|---|---|---|
| **AR(*p*)** | progressively decreases | Cuts off after *p* lags |
| **MA(*q*)** | Cuts off after *q* lags | progressively decreases |
| **ARMA(*p*,*q*)** | progressively decreases | progressively decreases |
| **ARIMA (*p*,*d*,*q*)** | Cuts off after *q* lags | Cuts off after *p* lags |

## 4.5. Quantile - Quantile plot

The quantile - quantile plot, often known as the Q-Q plot, is a graphical approach for determining if a set of data is likely to originate from any theoretical distribution, such as the normal or exponential distribution. A normal Q-Q plot, for example, may be used to test if our variable is normally distributed, after realizing a statistical analysis. It is a visual assessment that allows us to understand at a glance if our assumption is credible and, if not, how the assumption is violated and which data points contribute to the violation [8].

Q-Q plot takes your sample data and sort it ascendingly before plotting it against quantiles generated from a theoretical distribution. If both sets of quantiles originate from the same distribution, the dots should form (Figure 53) an almost straight line [7].

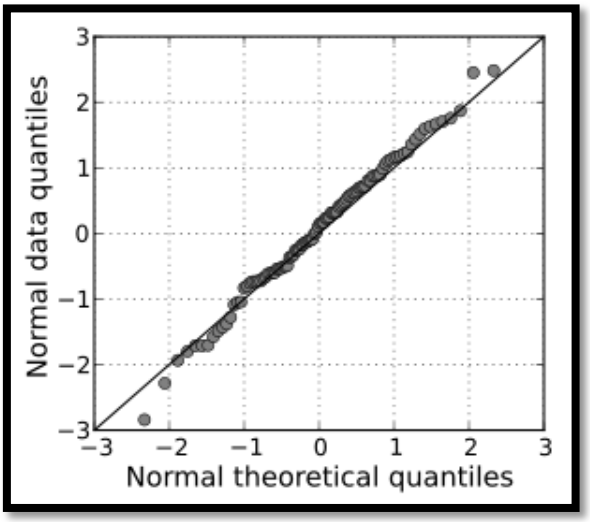



**Figure 53. An example of a normal distribution based on Q-Q plot [8].**

# 5. Methodology

The auto regressive integrated moving average (ARIMA) models developed by Box and Jenkin (1976) are the most widely used approach for time series analysis.

ARIMA models, developed by Box and Jenkins (Box *et al*., 2008), provide a robust approach to time series forecasting. Based on the autocorrelation analysis of the time series, a mathematical model to characterize the sequence may be developed (Chen *et al*., 2015).

Once the model is built, it is expected to predict future values considering the time series in the past and at present (Chen *et al*., 2015).

The assumption of some sort of statistical equilibrium is a key aspect in the creation of time series models (Hyndman and Athanasopoulos, 2013). Assumption of this kind is that of Stationarity (Pai and Lin, 2005). Forecasting is based on a linear combination of previous observations, which requires a stable time series with no discernible trend in the data (Pai and Lin, 2005). The ARIMA model assumes that the process remains at a statistical equilibrium with probabilistic properties that do not change over time, varying with a fixed constant mean level and with constant variance (Box *et al*., 2008). If the mean increases or decreases over time or if the variance does (as indicated by the excursions around the mean becoming smaller or larger over time), the series may need to be transformed to make it Stationary before modeling (Allard, 1998). Often, time series data are not Stationary, and they often have variations over time in means and variances (Helfenstein, 1991). Previously, researchers discovered that by differencing the time series, they could remove trend components in the mean (Helfenstein, 1991). Typically, one or two orders of differencing are sufficient to prepare data for the technique (Dindarloo, 2015). In addition to the trend components, time series related to the ecology field often show seasonal patterns. Box and Jenkins have developed a method to handle time series that contain seasonality (Box *et al*., 2008).

In this case, the model is known as SARIMA model with S observations per period. It is represented by SARIMA $(p,d,q)$ $(P,D,Q)_S$, which has the following form:

$$\varphi(B)\Phi(B^S)(1-B^S)^D\,(1-B)^d\,Z_t \;=\; \theta(B)\Theta(\mathrm{B}^S)_{\varepsilon t} \quad (1)$$

With

$$\varphi(B) = 1 - \varphi_1 B - \cdots - \varphi_p B^p \quad (2)$$

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q \quad (3)$$

$$\Phi\,(B^S) = 1 - \Phi_1 B^s - \cdots - \Phi_p B^{SP} \quad (4)$$

$$\Theta(\mathrm{B}^S) = 1 - \Theta_1 \mathrm{B}^S - \cdots - \Theta_Q \mathrm{B}^{SQ} \quad (5)$$

Where $p$ is the autoregressive order, $d$ is the number of differencing operations, $q$ is the moving average order and $P$, $D$, and $Q$ are the corresponding seasonal orders.

To build an ARIMA or SARIMA model for a time series, Box and Jenkins (1976) proposed an iterative approach (Tiao, 2001) involving four steps: Stationarity check, identification and estimation, diagnostic and residual check and prediction (Fig. 54).

This technique has become widely used, facilitating the practical application of time series models for forecasting. The Stationarity check phase involves suitable time series differencing. It is performed to achieve Stationarity and normality; the Stationarity of the time series is checked using the Augmented Dickey Fuller (ADF) test. The null hypothesis for this test is that the data is non-Stationary. For the data to be Stationary, we want to reject the null hypothesis, which means, a Stationary data with a p-value of equal or less than 0.05.

The temporal correlation structure of the transformed data (or not transformed, depending on the ADF test) is identified by examining its autocorrelation (ACF) and partial autocorrelation (PACF) functions (Mishra and Desai, 2005). The goal is to reduce the selection of parsimonious models worthy of further examination (Tiao, 2001). According to the ACF and PACF graphs of the Stationary series, various ARIMA models can be identified. ACF and

PACF plots are used to identify the six parameters (*p, d, q, P, D, Q*) in the SARIMA model. The minimum Aikake information criterion (AIC) and Bayesian information criterion (BIC) introduced by Akaike (1972) and Schwartz (1978), respectively, are used to select the best-fit model (Fraley and Raftery, 1998) among the candidate ones that were developed in the previous step.

AIC estimates the relative amount of information lost by a given model: the lesser information a model loses, the higher the quality of that model (Aho *et al*., 2014).

$AIC = \ln(\sigma_r{}^2) + 2r/T$

BIC is a criterion for model selection, which is closely related to the AIC. The model with the lowest BIC is preferred (Aho *et al*., 2014).

$BIC = \ln(\sigma_r{}^2) + rn(T)/T$

Where, $\sigma_r{}^2$ is the maximum likelihood estimation of the innovation variance, *r* is the number of parameters in the model and *T* is the size of the sample series.

After an appropriate model is chosen and the parameters are estimated, several tests are required (model fitness and residual checking) to verify whether the model is adequate for describing the studied process.

To do so, the ACF and quantile-to-quantile (Q-Q) figures of residuals were plotted. In addition, the forecast accuracy of the model was evaluated by splitting the data in two sets. The last observations in our data set were used to compare between simulated values and actual ones. Finally, the chosen SARIMA model was applied to predict the monthly Chl-*a* future values.

The long time series which covers approximately three decades of monthly Chl-*a* historical data, enabled us to perform multiple time steps ahead (Wei, 1990) from May 2018 to December 2025.

The modeling of the SARIMA model was performed using Econometrics toolbox in the MATLAB software MATLAB® software (version 7860349 (R2020b), The Mathworks, MA, USA).

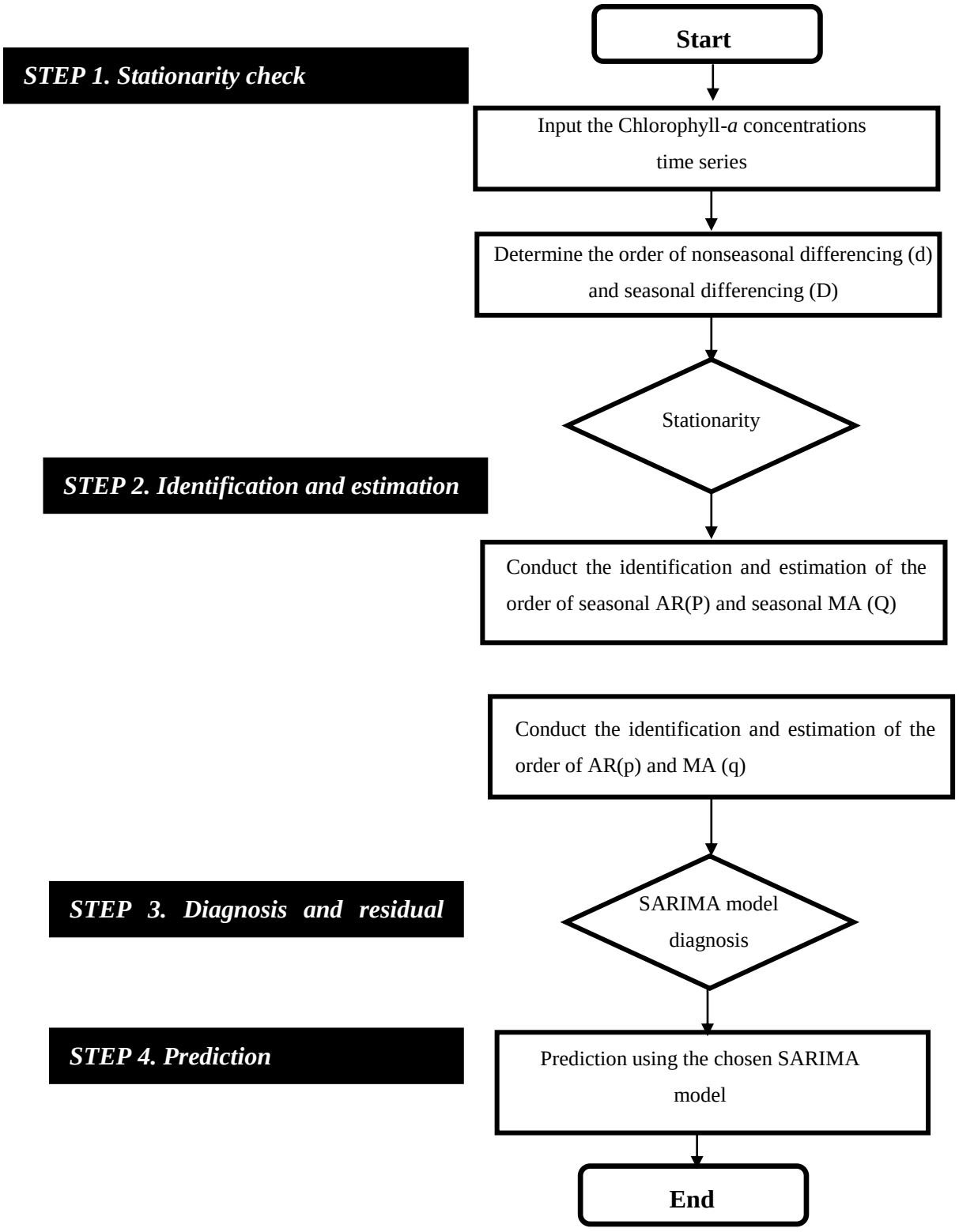



**Figure 54. The prediction process using the SARIMA model.**

## 6. Results and discussion

The time series of monthly Chl-*a* values is showed in figure 55. The autocorrelation pattern of Chl-*a* in the North Lagoon of Tunis (Fig.56) showed that the correlation coefficient declined relatively quickly. It was better to perform an ADF test to check the Stationarity of Chl-*a* concentrations time sequence (Table 8).

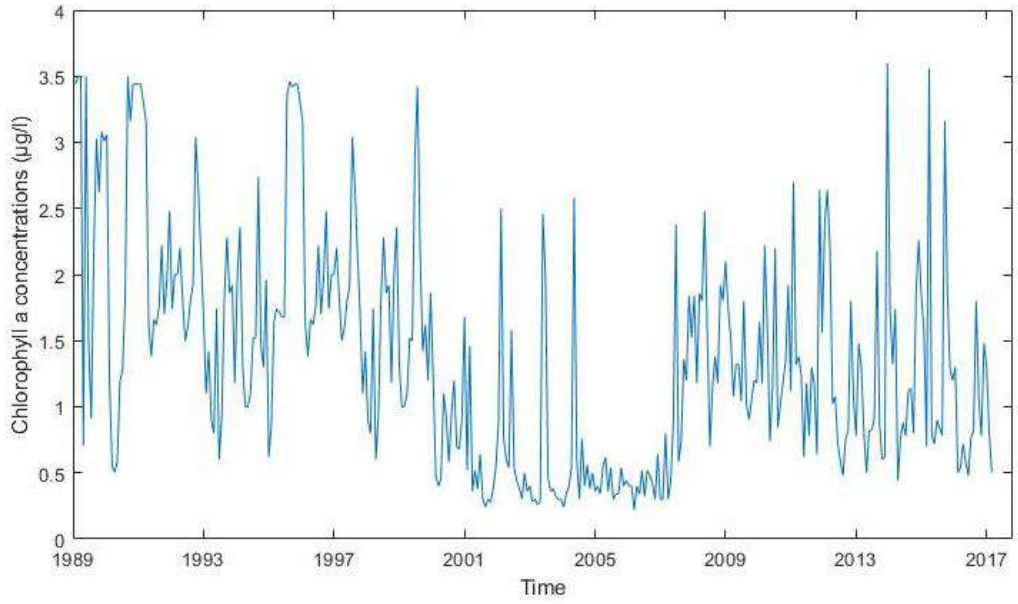



**Figure 55. Temporal distribution of Chlorophyll-*a* concentrations in the North lagoon of Tunis.**

The ADF test confirmed the Stationarity of Chl-*a* time series with a p-value < 0.05. According to that, the time series does not need transformation. It was not necessary to transform the data by differencing. Instead, we used the data on the original scale. Fitting a SARIMA model directly is advantageous for forecasting because forecasts are returned on the original scale.

**Table 10. ADF test results applied to the Chl-*a* original time series.**

| Null Rejected | P-value | Statistics Tests | Critical value | Significance level |
|---|---|---|---|---|
| **True** | 0.001 | -4.599 | -1.941 | 0.0500 |

Based on the PACF (Fig. 56), all SARIMA (*p, d, q, P, D, Q*) models in which the autocorrelation delay *p* and seasonal autocorrelation delay *P* was less than or equal to 4 and the moving average *q* and seasonal moving average *Q* was less than or equal to 4, were tested. As mentioned above, Chl-*a* time series data contain a seasonal component. To analyze its time series from January 1989 through April 2018, we defined S=12 because we have 12 observations per year.

It was found the minimal AIC and BIC information when performing a SARIMA $(2,0,2)(2,0,2)_{12}$ with AIC = 628.91, BIC = 666.78 and $R^2 = 0.52$. To avoid making remarkable changes in the original data, it is better to keep the number of parameters to a minimum, so that the values of p, P, q, Q, d, and D selected are less than or equal to 2 (Hintze, 2007).

It was observed that our developed SARIMA $(2,0,2)(2,0,2)_{12}$ for Chl-*a* forecasting with AIC= 628.91 and BIC=666.78 and $R^2$=0.52 performed better than SARIMA$(1,0,0)(2,0,0)_{12}$ with AIC = 1593, BIC =1612 and $R^2$= 0.46 reported by Raman *et al*. (2018) for forecasting monthly fish catch data in Chilika lagoon situated in East Coast of India. In addition, our SARIMA $(2,0,2)(2,0,2)_{12}$ was better than the SARIMAX with AIC = 1114.2, BIC = 1141.0, and $R^2$ = 0.52 developed also by Raman *et al*. (2018). Even though, the coefficient of determination $R^2$ is similar, our model presented the lowest AIC and BIC.

It is to be mentioned that the SARIMAX, is a model developed using Box and Jenkins (1976). SARIMAX is a SARIMA with external factors derived using PCA analysis (Raman *et al*.; 2018).

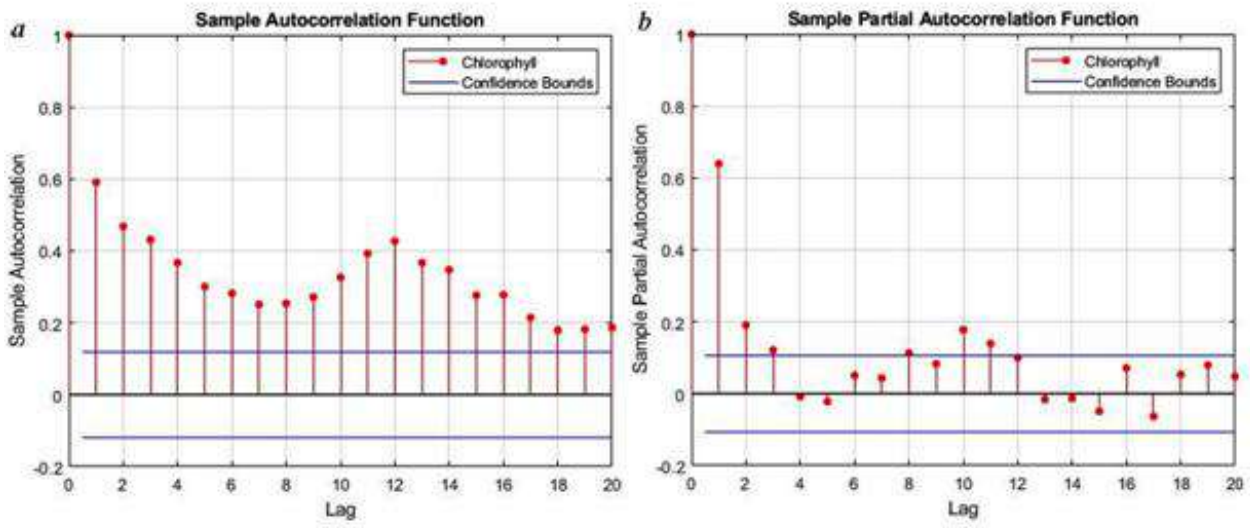



**Figure 56. Autocorrelation Function (a) and Partial Autocorrelation Function (b) of the monthly Chlorophyll-*a* variations in the North lagoon of Tunis.**

Estimation of the SARIMA $(2,0,2)(2,0,2)_{12}$ model parameters and their testing results are presented in table 9. All estimated coefficients are statistically significant (p-value < 0.05).

**Table 11. Parameter estimates and their testing results of the SARIMA $(2,0,2)(2,0,2)_{12}$ model.**

| Parameter | Coefficient | Standard Error | T statistics | p-value |
|---|---|---|---|---|
| **Constant** | 0.0448 | 0.0031 | 2.4357 | 0.01511 |
| **AR (1)** | 0.0896 | 0.0400 | 1.9890 | 0.03227 |
| **AR (2)** | 0.0812 | 0.0333 | 2.2383 | 0.02156 |
| **SAR (12)** | 0.0967 | 0.0435 | 2.2250 | 0.02156 |
| **SAR (24)** | 0.6763 | 0.0342 | 19.7782 | 4.5892e-87 |
| **MA (1)** | -0.0686 | 0.3935 | -0.0217 | 0.0382 |
| **MA (2)** | -0.3391 | 0.1912 | -1.7739 | 0.0491 |
| **SMA (12)** | -0.0983 | 0.0577 | 1.7045 | 0.0488 |
| **SMA (24)** | -0.6917 | 0.0578 | -11.6159 | 5.3588e-33 |
| **Variance** | 0.3302 | 0.0225 | 14.6879 | 7.7121e-41 |

The model fit analysis is done by residuals checking. The residuals ACF of SARIMA $(2,0,2)(2,0,2)_{12}$ is presented in Figure 57. Residuals ACF suggested autocorrelations near 0. This indicates that the residuals did not deviate significantly from a 0 mean. In more plain words, it means that the residuals are not correlated.

The residual Q-Q plot (Fig.58) suggests that the residuals are approximately normally distributed, with slightly heavier tail. Given that, we can conclude that, residuals of the SARIMA $(2,0,2)(2,0,2)_{12}$ model are uncorrelated and approximately normally distributed.

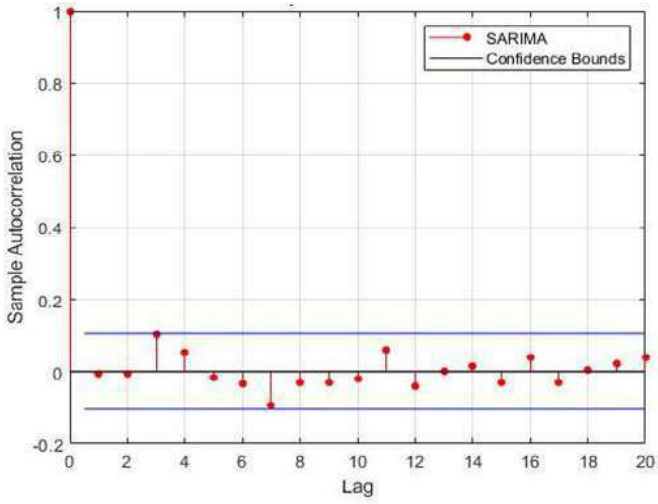



**Figure 57. Autocorrelation Function (ACF) of residuals.**

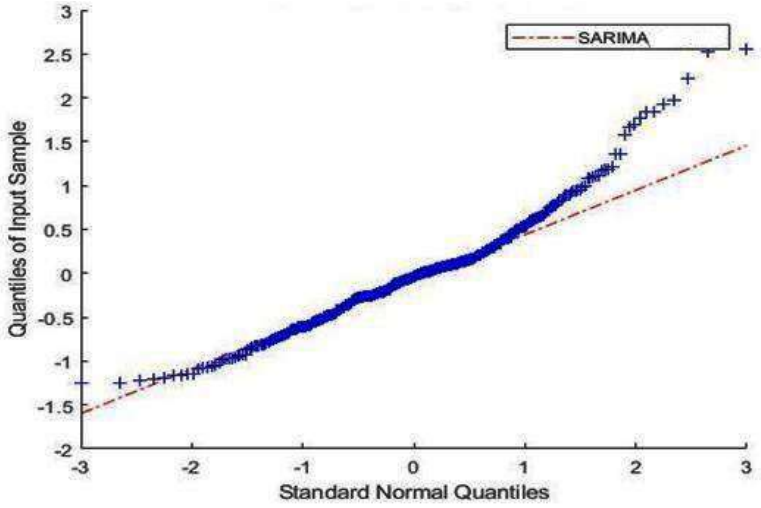



**Figure 58. Residual Quantile - Quantile (Q-Q) plot.**

Comparison of observed Chl-*a* data with the fitted ones by the SARIMA $(2,0,2)(2,0,2)_{12}$ model are presented in figure 59. There is a good match between the observed values and the fitted ones. To check the forecasting accuracy, table 10, shows a comparison between the predicted values and the observed ones for the period from January 2017 to April 2018.

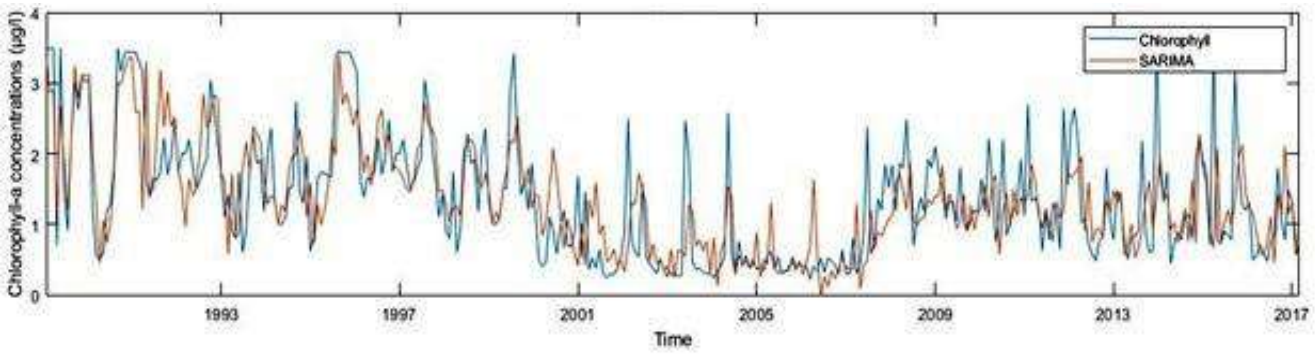



**Figure 59. SARIMA $(2,0,2)(2,0,2)_{12}$ model fit of the Chlorophyll-*a* content time series in the North lagoon of Tunis from January 1989 to April 2018.**

**Table 12. Comparison of predicted and observed monthly Chlorophyll-*a* variations using the SARIMA $(2,0,2)(2,0,2)_{12}$ model for the data from January 2017 to April 2018.**

| Month | Predicted Values | Observed Values |
|---|---|---|
| **January** | 1.44 | 1.2 |
| **February** | 1.22 | 1.3 |
| **March** | 1.15 | 0.5 |
| **April** | 0.84 | 0.54 |
| **May** | 0.68 | 0.72 |
| **June** | 0.74 | 0.58 |
| **July** | 0.57 | 0.48 |
| **August** | 0.91 | 0.76 |
| **September** | 0.74 | 0.82 |
| **October** | 1.52 | 1.8 |
| **November** | 0.87 | 1.06 |
| **December** | 0.62 | 0.78 |
| **January** | 1.23 | 1.48 |
| **February** | 0.98 | 1.3 |
| **March** | 0.76 | 0.80 |
| **April** | 0.63 | 0.5 |

In this study, we have applied a SARIMA $(p,d,q)(P,D,Q)_S$ model to analyze Chl-*a* variations of monthly collected data in the North lagoon of Tunis from 1989 through 2018, in the purpose of assessing and forecasting eutrophication and contributing to prevent any deterioration in the studied ecosystem. We have developed a SARIMA that closely fitted Chl-*a* observed data. According to our results, the SARIMA model developed in this study was reliable with high validity, which suggests that our model could be an appropriate statistical tool to predict the future changing trends of Chl-*a* values or any other key parameters, thus preventing high eutrophication scenarios in the North lagoon of Tunis or other ecosystems.

SARIMA has several advantages over other predictive methods, such as moving average, exponential smoothing, etc, in particular, because of its forecasting capability, especially in case of long periods of time, and its richer information on time-related changes (Linthicum *et al*., 1999; Box *et al*., 2008).

Once we obtained the satisfactory model, we have used it to forecast future values of the Chl-*a* in the ecosystem. Figure 60 shows the forecast of Chl-*a* concentrations in the North lagoon of Tunis. The model seems to provide realistic predictions in the future.

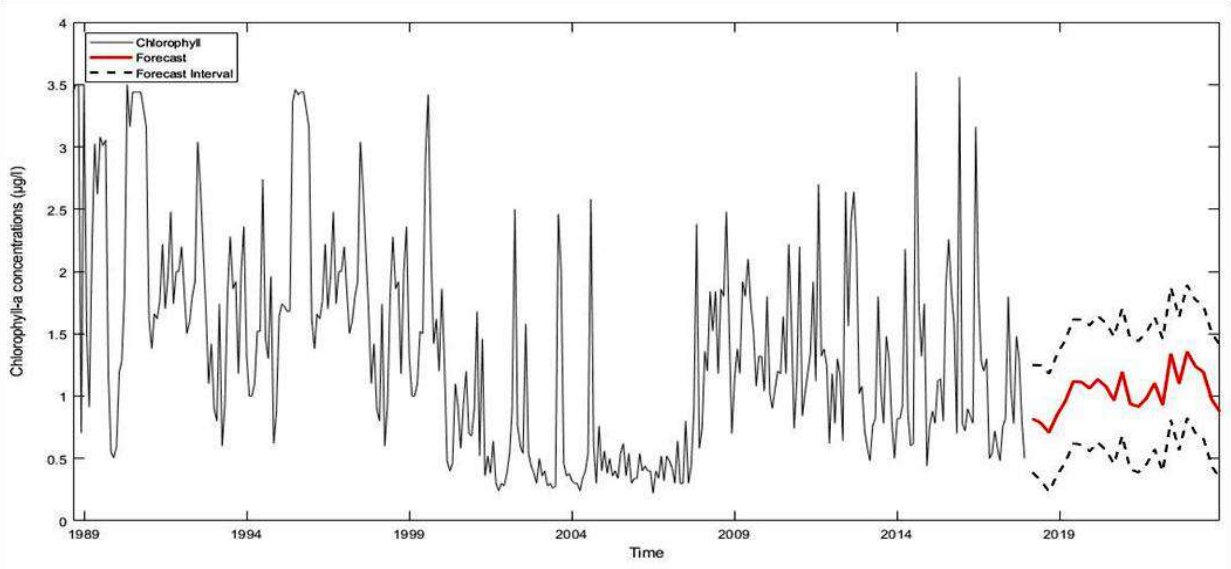



**Figure 60. Time series plot of Chlorophyll-*a* concentrations in the North lagoon of Tunis with forecasts and forecast intervals at 75%.**

To the best of our knowledge, this is the first study that has applied the SARIMA model to fit and forecast monthly Chl-*a* variations, in the North lagoon of Tunis.

Forecasting plays an essential role in policy formulation and implementation especially in the management of fisheries or aquatic resources. In this context, several time series model have been developed for forecasting purposes, such as ARIMA model, multiple linear regression (MLR), non-linear regression (NLR), smoothing models, dynamic models, harmonic regression (HREG), Vector auto regression model, generalized auto regressive conditional heteroscedasticity (GARCH), Gausian autoregressive models (Stergiou 1991; Stergiou *et al*.

1997; Romilly 2005), the studies showed that ARIMA validation errors are significantly lower, and is a better forecasting model (Raman *et al*., 2018).

It is known that the SARIMA modeling process needs a large amount of data. An early study suggested that a minimum of 50 observations are needed to build a reasonable SARIMA model (Wei, 1990), explaining most of its variance and adequately modeling the seasonality and correlation structure of the data (Stergiou 1991; Pajuelo and Lorenzo 1995). To obtain a stable and precise SARIMA model, we had to collect 352 observations of Chl-*a* data over the past three decades without interruption to be mentioned. We can state that the results of our study are robust enough.

Similar, ARIMA model has been used in various studies in great lagoons such as on net basin supply (NBS) for extreme flood study in lakes Eric and Ontario (Mathier *et al*. 1992), on modeling of the great lakes freeze for ferrous scrap (Albertson and Aylen 1996), and on spatio-temporal behavior of atmospheric temperature data of great lakes US (Agrawal 2011) and reliable predictions by ARIMA with acceptable accuracy showed a important role for aquatic natural ecosystems managers.

The SARIMA model may not have a good interpretation of the eutrophication process or internal/external factors affecting Chl-*a* rates, but from implementation point of view it is the simplest and most suitable model to apply over other nonlinear models (Adhikari and Agrawal 2013). Indeed, the model needs only one input variable and provides predictions with acceptable accuracy, which makes its applicability easier for practical applications in early-warning of algal blooms. In addition, it is important to mention that working with only one variable is money and time-saving for researchers.

## 7. Station 5

For comparison, the Box en Jenkins methodology was realized on the observed Chl-*a* data of the Station 5. Indeed, this area of the North Lagoon of Tunis was identified the most influenced by the pollution, regarding to its position in the ecosystem.

An ADF test was performed to check the Stationarity of the Chl-*a* time series at the Station 5 of the lagoon (p-value = 0.1). The ADF test confirmed the non-Stationarity of the Chl-*a* time series with a p-value > 0.05.

According to that, it was necessary to transform the data by differencing. It's possible that this finding is due to the data in Station 5 being more heterogeneous and having more increased nonlinear interactions.

Based on the autocorrelation pattern of the Chl-*a*, ACF and PACF (Fig.61), all SARIMA ($p, d, q, P, D, Q)_S$ models in which the autocorrelation delay $p$ and seasonal autocorrelation delay $P$ was less than or equal to 3 and the moving average $q$ and seasonal moving average $Q$ was less than or equal to 3, were tested.

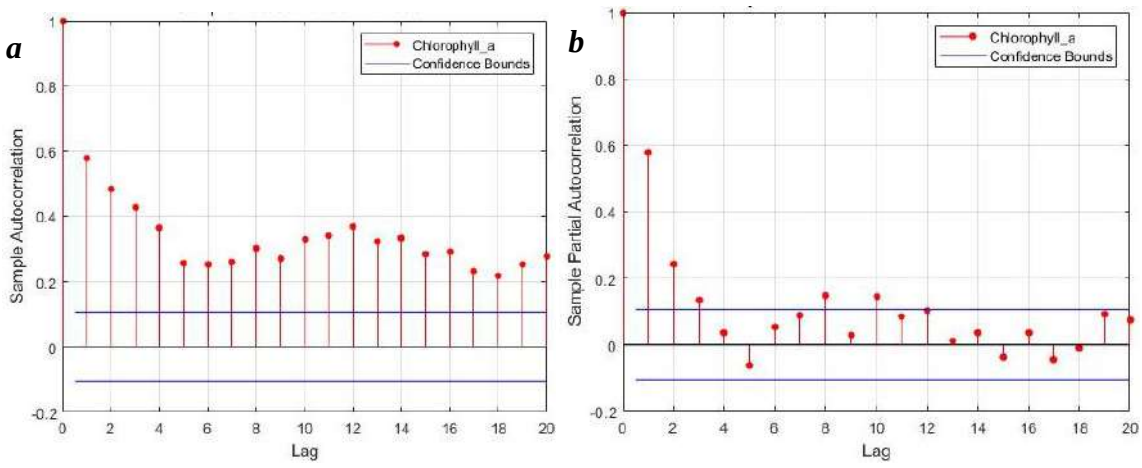



**Figure 61. ACF (a) and PACF (b) of the monthly Chl-*a* variations in the Station 5.**

It was found the minimal AIC and BIC information when performing a SARIMA $(1,1,0)(0,1,0)_{12}$, with AIC = 1210.0 and BIC = 1222.2.

The performance of the SARIMA slightly decreased in comparison with the SARIMA $(2,0,2)(2,0,2)_{12}$ realized with monthly dynamics of the Chl-*a* in the lagoon, given the heterogeneity of the data at Station 5.

The differences between observed and simulated data from Station 5 are clearly less significant than the differences between observed data from Station 5 and simulated data from the entire lagoon performed earlier in this manuscript (Figure 62). Given this, it is critical to create the SARIMA model for Station 5 and closely monitor the Chl-*a* levels in this part of the lagoon.

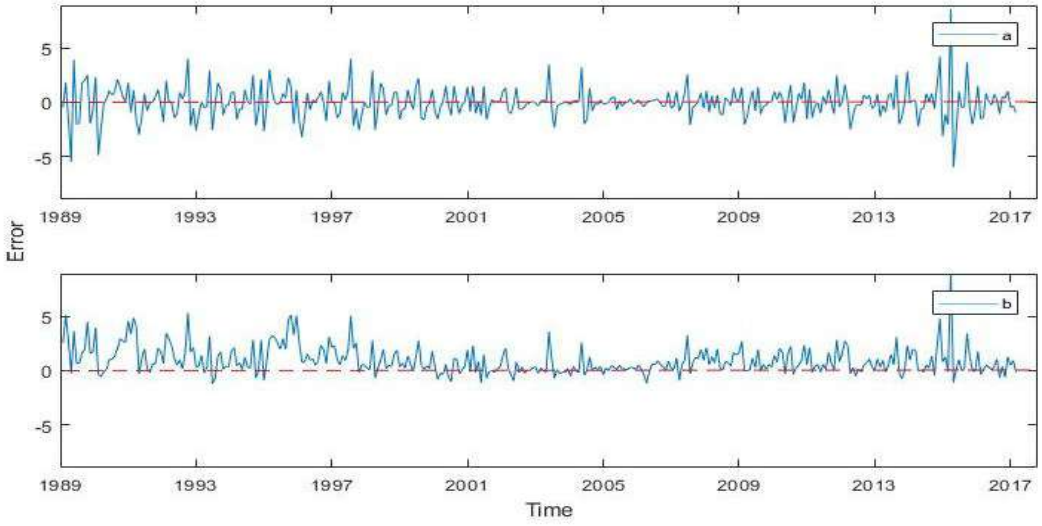



**Figure 62. (a) Variations between observed Station 5 Chl-*a* data and simulated Station 5 Chl-*a* data (mean = 0.021 +/- 0.002), and (b) variations between observed Station 5 Chl-*a* data and simulated Chl-*a* data from the lagoon data (mean = 0.95 +/- 0.01).**

The ACF of SARIMA $(1,1,0)(0,1,0)_{12}$ residuals is presented in Figure 63. Residuals did not deviate considerably from a 0 mean, which means that the residuals are uncorrelated.

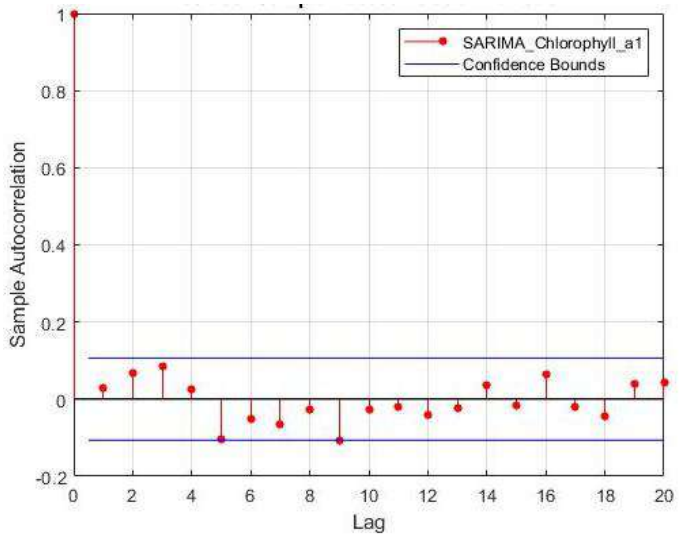



**Figure 63. ACF of residuals.**

The residual Q-Q plot (Fig. 64) suggests that the residuals are approximately normally distributed, with slightly heavier tail. Given that, we can conclude that, residuals of the SARIMA $(1,1,0)(0,1,0)_{12}$ model are uncorrelated and normally distributed.

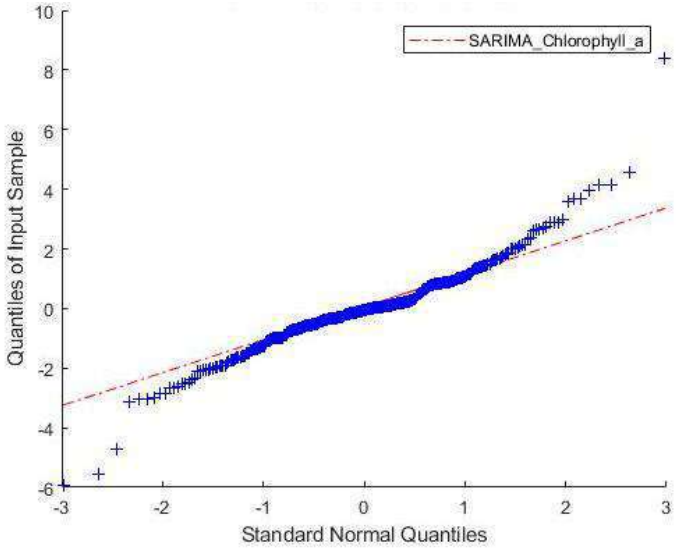



**Figure 64. Residuals (Q-Q) plot.**

Figure 65 compares the observed Chl-*a* data with the model-fitted data using the SARIMA $(1,1,0)(0,1,0)_{12}$ model. We can see that between the measured values and the fitted values, there is a good agreement.

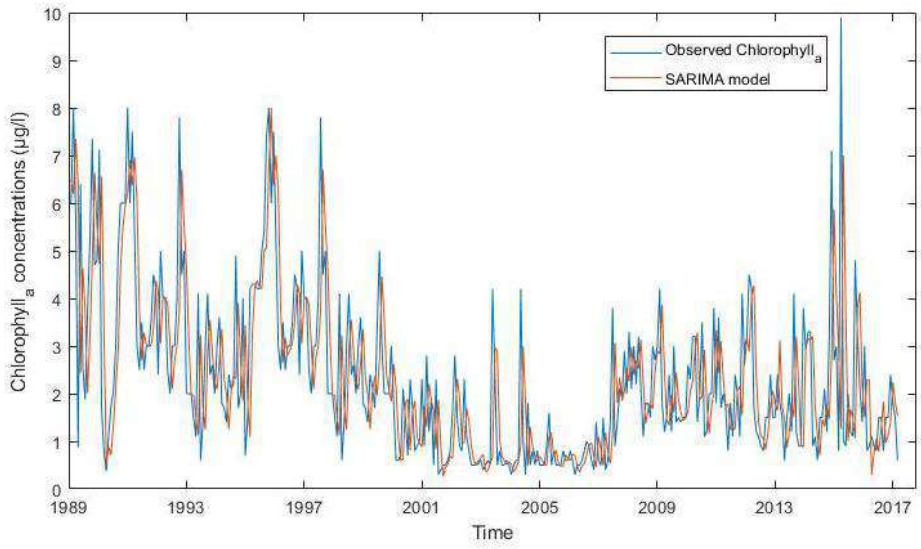



**Figure 65. SARIMA (1,1, 0) $(0,1,0)_{12}$ model fit of Chl-*a* content time series in Station 5.**

Once we had a good model, we were able to predict future Chl-*a* values in the ecosystem at Station 5. Figure 66 represents the predicted Chl-*a* concentrations in Station 5 from May 2018 to December 2025. The model appears to be capable of making accurate forecasts for the future, despite the fact that the model underestimated the high values of the Chl-*a*.

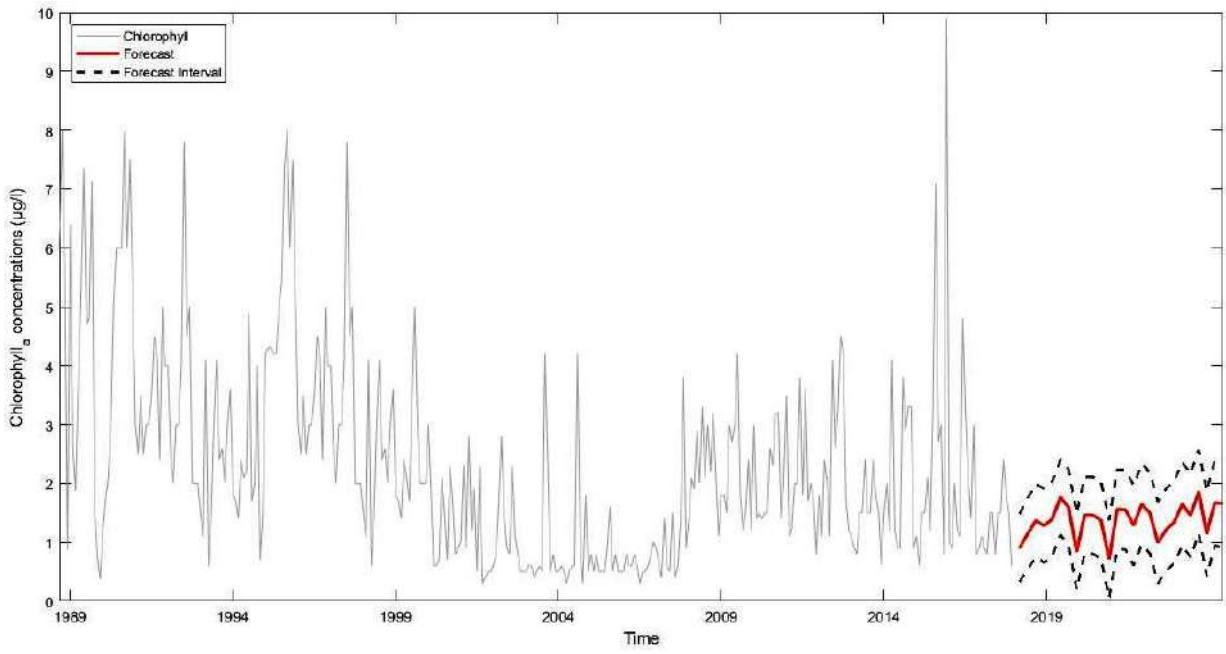



**Figure 66. Time series plot of Chl-*a* concentrations in the North Lagoon of Tunis Station 5 with forecasts and at 75% confidence interval.**

The data sets underestimation is most likely owing to the non-homogeneous nature of the data water quality characteristics. These data were gathered over a 30-year period, resulting in a significant variation in the values of water quality indicators in this study, especially at Station 5.

A comparison is conducted between model results simulated using monthly dynamics of Chl-*a* in the whole lagoon and model results simulated only using data from Station 5 to assess the spatial heterogeneity in the lagoon.

It is shown that with only one input variable , the SARIMA created was able to forecast Chl-*a* concentration dynamics quite effectively in Station 5.

As mentioned earlier in this manuscript, the North Lagoon of Tunis offers excellent fishing opportunities. Fisheries are installed in the south-east zone of the ecosystem, to catch fishes how are trying to reach the open sea. As a result, controlling water quality in this section of the lagoon is critical.

## 8. Conclusion

In our study, we have used the SARIMA model to forecast future values of Chl-*a* in the North Lagoon of Tunis as an eutrophication indicator. This model was applied for a time window of approximately three decades. Different SARIMA (p,d,q)$(P,D,Q)_S$ models were implemented. The chosen one, a SARIMA $(2,0,2)(2,0,2)_{12}$ with the lowest AIC and BIC was used for forecasting.

The goodness of fit was analyzed by comparing with the actual data from last observations and checking the residuals. The residual diagnostic indicated that they are uncorrelated and approximately normally distributed.

The forecasting results are quite satisfactory (AIC=628.91; BIC=666.78 and $R^2$=0.52) since the forecasting period seems to reproduce relatively well the normal Chl-*a* monthly content in the lagoon.

The same approach was simulated on historical data of the Station 5, which seems to be the most affected area with eutrophication in the lagoon. The created SARIMA $(1,1, 0)\ (0,1,0)_{12}$ for Station 5, was able to forecast the Chl-*a* concentration quite effectively.

Despite the fact that the North lagoon of Tunis was classified as a Wetland of International Importance (Ramsar site), given its history and importance, this ecosystem is still a fragile one. The SARIMA model applied to historical data of Chl-*a* or any other key parameter could be an important tool to providing early evidence that guides prevention and control interventions for the ecosystem.

# Chapter IV: Hybrid model: Integrating artificial intelligence and Box and Jenkins time series models (SARIMA-NARX)

## 1. Introduction

One of today's most challenging environmental issues impacting surface water systems is eutrophication (Li *et al*., 2017). Hydro-climatic variables such as evaporation, temperature, precipitation, etc. in addition to anthropogenic contribution, made eutrophication qualified as a complicated process to control (Schramm, 1999; Viaroli *et al*., 2008).

Proper ecosystem management cannot be achieved without access to precise data to realize accurate predictions of eutrophication episodes for effective resource conservation operations (Borja *et al*., 2008). Forecasting, as a form of early monitoring and pollution detection, can help in the creation of successful environmental control measures (Tian *et al*., 2017).

Accurate forecasting of eutrophication episodes is critical for more timely and efficient ecosystem management (Borja *et al*., 2008). If natural resources are managed in a sustainable manner, coastal ecosystems should be able to supply goods and services that support varied human needs (Borja *et al*., 2008).

Previous research principally implemented two methods for time series forecasting of eutrophication, using the Chlorophyll-*a* as an eutrophication indicator: (i) linear approach: MVLR (multivariate linear regression); Box and Jenkins techniques (ARIMA, SARIMA,…); and (ii) non-linear approach: ANNs (artificial neural networks). First, the linear method is useful when the independent and dependent variables have a linear relationship (Koo *et al*., 2010; Hong *et al*., 2012).

The ARIMA model, created by Box and Jenkins, uses auto-correlation analysis of time series data to predict future values (Box and Jenkins, 1976). The SARIMA model can be utilized if there is a seasonal trend in time series data (Zhang *et al*., 2003).

Sun and Koch (2001) predicted the future values of water salinity in Apalachicola bay in Florida using an ARIMA model. Hassanzadeh *et al*. (2009) and Makwinja *et al*. (2017) used ARIMA model to forecast sulphur dioxide in Tehran and water levels in Lake Malawi, respectively.

When compared to exponential smoothing and to MVLR model, the ARIMA model was shown to be the most superior (Jeong *et al*., 2014).

Second, the non-linear method is appropriate in cases when the independent and dependent variables have a non-linear relationship (Jeong *et al*., 2014).

Previous researches have used ANN models in non-linear methods as a representative technique. Wang *et al*. (2010) predicted cyanobacterial blooms in Dianchi Lake in China based on weather conditions using ANN methods. Li *et al*. (2017) and Yi *et al*. (2018) applied different types of artificial neural networks to estimate the concentration of Chl-*a* in 27 lakes in China and in one Korean river, respectively.

Compared to the MVLR model and to support vector machine, the ANN model was shown as the most superior (Jeong *et al*., 2014).

Zhang and Qi (2005) created an ANN model that takes seasonality into account, emphasizing the need of appropriate data pre-processing to account for seasonal or trend fluctuation.

Real processes, often have both linear and nonlinear characteristics, and earlier research tried to clarify the variability of real problems by combining two techniques (Chen *et al*., 2007). For example, coastal lagoons are vulnerable to regular environmental disturbances and variations (evaporation, precipitation, etc.), in addition to a tight link with temperature and season dependency. As a result, the combined model, often known as the hybrid model, has become a popular technique for improving forecasting accuracy.

In neural network forecasting research, several combining techniques have been studied. Haizum Abd Rahman *et al*. (2019) described a combining methodology using Artificial

Neural Network and the Seasonal ARIMA to forecast air pollutant index in Johor city, Malaysia. Yu *et al*. (2019) presented a hybrid SARIMA and NARX network approach for predicting the incidence of schistosomiasis in China. Pelikan *et al*. (1992) and Ginzburg and Horn (1994) proposed to combine several feed-forward neural networks to improve time series forecasting accuracy.

In the previous chapters Seasonal Autoregressive integrated moving average (SARIMA) model and artificial neural network NARX model have performed well in forecasting linear and non-linear Chl-*a* content time series in the North Lagoon of Tunis, individually as an eutrophication indicator.

The aim of this chapter is to create a predictive model for forecasting Chl-*a* content in the North Lagoon of Tunis by combining the SARIMA model (linear method) with the NARX network (non-linear approach).

This chapter was conducted in three steps: (1) establishment of Chl-*a* concentrations data in the North Lagoon of Tunis; (2) Forecasting the monthly Chl-*a* variation using the linear approach (SARIMA model); and (3) improvement in prediction accuracy by considering the non-linearity of the residual from the SARIMA model using the NARX model.

## 2. Time series forecasting models

Time series modeling may be done in a variety of approaches. One of the most appropriate and often used time series technique is the autoregressive integrated moving average (ARIMA) model (Zhang, 2003).

The popularity of the ARIMA model stems from its statistical characteristics throughout the model-building phase (Box and Jenkins, 1970). One of the basic assumptions of the ARIMA technique, and hence of traditional prediction methodologies, is that the time series under consideration is the result of linear processes (Zhang, 2003). The ARIMA model implies that future values are linearly related to present and previous time series values. Nevertheless, because real-world processes are complicated and frequently dynamic and non-linear (Zhang *et al*.,1998), traditional techniques may be insufficient if the time series exhibits non-linear behavior (Zhang, 2003).

Many nonlinear models have been presented as alternate techniques to solve the problem of nonlinearity, with artificial neural networks (ANNs) being one of the most frequent and essential approaches. The capacity of neural networks to model nonlinearly is its primary benefit. ANN models can approximate a nonlinear process of any complexity (Hornik *et al*.,1989). Given these characteristics, it is not unexpected that ANNs have piqued the interest of time series forecasters (Zhang, 2003).

## 3. Hybrid model

The arguments illustrating the necessity for hybrid model are highlighted in this section.

First, it is sometimes difficult in practice to determine if a time series under investigation is created by a linear or nonlinear underlying mechanism, or whether one technique is more efficient than the other. As a result, forecasters have difficulties in determining the appropriate

technique for their specific scenarios. Several models are often tested, and the one that produces the most accurate results is chosen (Zhang, 2003).

Second, time series in the actual world are rarely pure linear or nonlinear. They frequently include both linear and nonlinear patterns. If this is the case, neither ARIMA nor ANNs can be used to model and predict time series since the ARIMA model cannot handle nonlinear correlations and the neural network model cannot handle both linear and nonlinear patterns equally effectively. As a result, by merging ARIMA and ANN models, complicated autocorrelation structures in data may be more precisely represented (Zahng, 2003).

Third, it is nearly widely acknowledged in the forecasting literature that no single technique is superior in any case (Jenkins, 1982; Makridakis *et al*., 1982; Chatfield, 1988). This is largely due to the fact that a real-world situation is always dynamic, and different patterns cannot be described equally effectively by a single model (Khandelwal *et al*., 2015).

For example, in the literature of time series forecasting with neural networks, most studies (Sharda *et al*., 1990; Tang *et al*., 1991; Tang *et al*., 1993; Zhang *et al*., 2001; Hwang *et al*., 2001) evaluate the efficiency of the ANN model using the ARIMA models as a reference.

Many studies, show that combining multiple distinct models may often increase forecasting accuracy over the individual model without the requirement to select the "best" model (Newbold and Granger, 1974; Makridakis *et al*., 1982; Clemen, 1989; Makridakis *et al*., 1993). As a result, combining several models can improve the capacity to capture various data patterns and improve performance predicting. Using a hybrid model or integrating different models to increase forecasting accuracy has become widespread practice (Makridakis *et al*., 1982). In forecasting, the primary principle behind model combination is to use each model's distinct feature to catch diverse patterns in the data (Zhang, 2003).

# 4. Methodology

## 4.1. Hybrid model

As previously stated, neither ARIMA nor ANN are universally applicable to all forms of time series. This is due to the fact that all real-world time series contain both linear and nonlinear correlation patterns between observations.

The traditional approach of the hybrid ARIMA–ANN model introduced by Zhang, (2003) considers that time series can be decomposed as the sum of a linear component and a nonlinear component. Then,

$$Y_t = L_t + N_t \tag{9}$$

Where $L_t$ denotes the linear component and $N_t$ denotes the nonlinear component to be estimated.

The ARIMA model is used in the first stage to fit the linear component and, as a consequence, to produce the prediction value denoted as $\widehat{L_t}$. Over-fitting, which is more closely associated with neural network models, can be eased by first fitting the ARIMA model to the data (Zhang, 2003).

The residual at time t can be obtained by comparing the real value ($Y_t$) with the predicted value ($\widehat{L_t}$). That is,

$$e_t = Y_t - \hat{L}_t \tag{10}$$

Because the residuals dataset after ARIMA fitting contains only nonlinear components, it can be appropriately simulated using an ANN.

With $n$ input nodes, the ANN for residuals has the following form:

$$e_t = f(e_{t-1}, e_{t-2}, \dots, e_{t-n}) + \varepsilon_t \tag{12}$$

Where $f$ is a nonlinear function determined by the neural network and $\varepsilon_t$ is the random error. If $(\widehat{N_t})$ is the forecast of this ANN, then the ultimate hybrid forecast at time *t* is obtained as:

$$\widehat{Y_t} = \widehat{L_t} + \widehat{N_t} \quad (13)$$

The proposed methodology in this chapter for forecasting the content of Chl-*a* as a eutrophication indicator in North lagoon of Tunis consists in the development of a hybrid model that combines a SARIMA model for the seasonal and linear characteristics of the Chl-*a* time series and a NARX model as a neural network method for the nonlinear characteristics of the time series.

For Chl-*a* forecasting, the SARIMA $(2,0,2)(2,0,2)_{12}$ model, which was shown to be the most appropriate in the previous chapter, is utilized primarily. The residuals of the SARIMA model, along with Secchi depth and dissolved oxygen data, are then simulated using the NARX neural network to capture the time series' nonlinearity.

Chapter II in section (3.3.1) and Chapter III in section (6) provide the followed methodology in developing the NARX network and the SARIMA $(2,0,2)(2,0,2)_{12}$ model, respectively.

After being satisfied with the performances of the two models separately, we can then combine the two approaches to obtain the performed forecasts (Fig. 67).

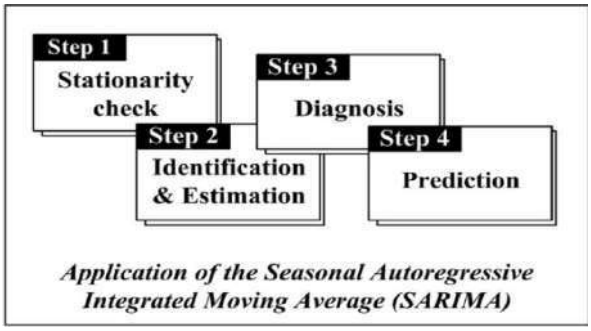
***STEP 1. Establishment of Chlorophyll-a concentrations data in the North Lagoon of Tunis***

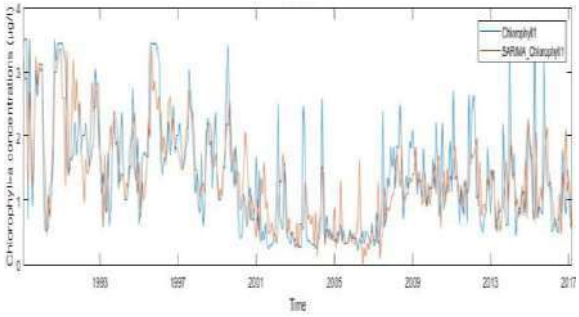


*Analysis of the physico-chemical and biological characteristics of the North Lagoon of Tunis*

*Monthly Chlorophyll-a and physico-chemical collected data from the North Lagoon of Tunis*

***STEP 2. Prediction of the monthly collected Chlorophyll-a data in the North Lagoon of Tunis using the SARIMA model***

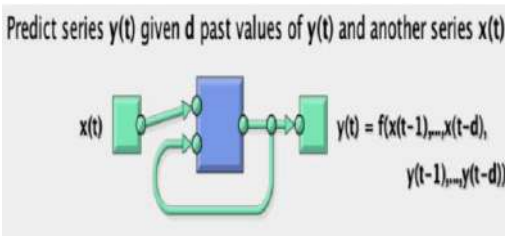


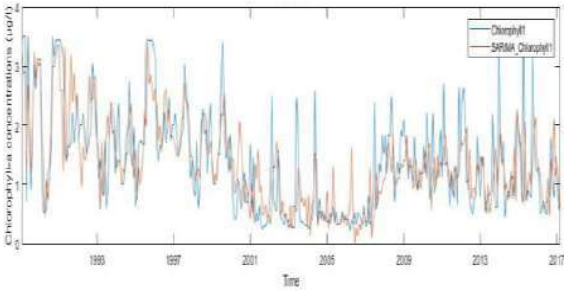




*Application of the Seasonal Autoregressive Integrated Moving Average (SARIMA)*

*Estimation of the monthly Chlorophyll-a content in the North Lagoon of Tunis*

***STEP 3. Improvement of the prediction accuracy by considering the non-linearity of the residual from the SARIMA model using the NARX network model***




*Application of the non linear Neural Network with Exogenous Input (NARX)*

*Improvement of the prediction accuracy*

**Figure 67. Research framework.**

## 4.2. Performances measures

Several criteria have been examined in order to generate reliable results for generating predictions and evaluating the performance of the proposed model. The ultimate goal is to ensure that the best model is chosen.

The estimation of generalization error has been calculated using three performance indexes: the standard correlation coefficient (R), the mean square error (MSE) and the coefficient of determination ($R^2$). Section (3.2) in Chapter II presents these performance metrics and their calculations. It is well recognized, that when comparing and selecting an appropriate model from a large number of options, it is advisable to employ as few mathematical measurements as possible [4].

In this chapter, all NARX modeling were performed using the MATLAB software MATLAB® software (version 9.3.0.948333 (R2017b), The Mathworks, MA, USA).

The modeling of the SARIMA model was performed using Econometrics toolbox in the MATLAB software MATLAB® software (version 7860349 (R2020b), The Mathworks, MA, USA).

# 5. Results and discussion

A SARIMA $(2,0,2)(2,0,2)_{12}$ has been found to be the most parsimonious among all SARIMA models tested on the historical Chl-*a* data in the North Lagoon of Tunis, according to the AIC and BIC, also, judged by the residual analysis.

The neural network model used is a NARX network as also employed by Benihabib *et al*. (2017) and yu *et al*. (2019) in developing hybrid approaches.

Table 11 gives the forecasting results for the Chl-*a* concentrations in the lagoon using the different approaches. Results show that applying NARX alone can improve the forecasting accuracy over the SARIMA model. Nevertheless, it seems that the performance of the NARX

decreases, especially during the last 5 years of the study period, at some points of high values of Chl-*a*.

Figure 68 gives the actual vs. simulated values with individual models of NARX and SARIMA as well as the combined model.

This may suggest that neither the neural network nor the SARIMA model captures all of the specificities in the data.

**Table 13. Forecasting comparison for Chlorophyll-*a* concentrations using the different modeling approaches.**

| | MSE | R | $R^2$ |
|---|---|---|---|
| NARX network | 0.31 | 0.79 | 0.62 |
| SARIMA model | 0.37 | 0.72 | 0.52 |
| Hybrid model | 0.24 | 0.82 | 0.67 |

The results of the hybrid model show that by combining two models together, the overall forecasting errors can be significantly reduced compared with the SARIMA model and NARX network.

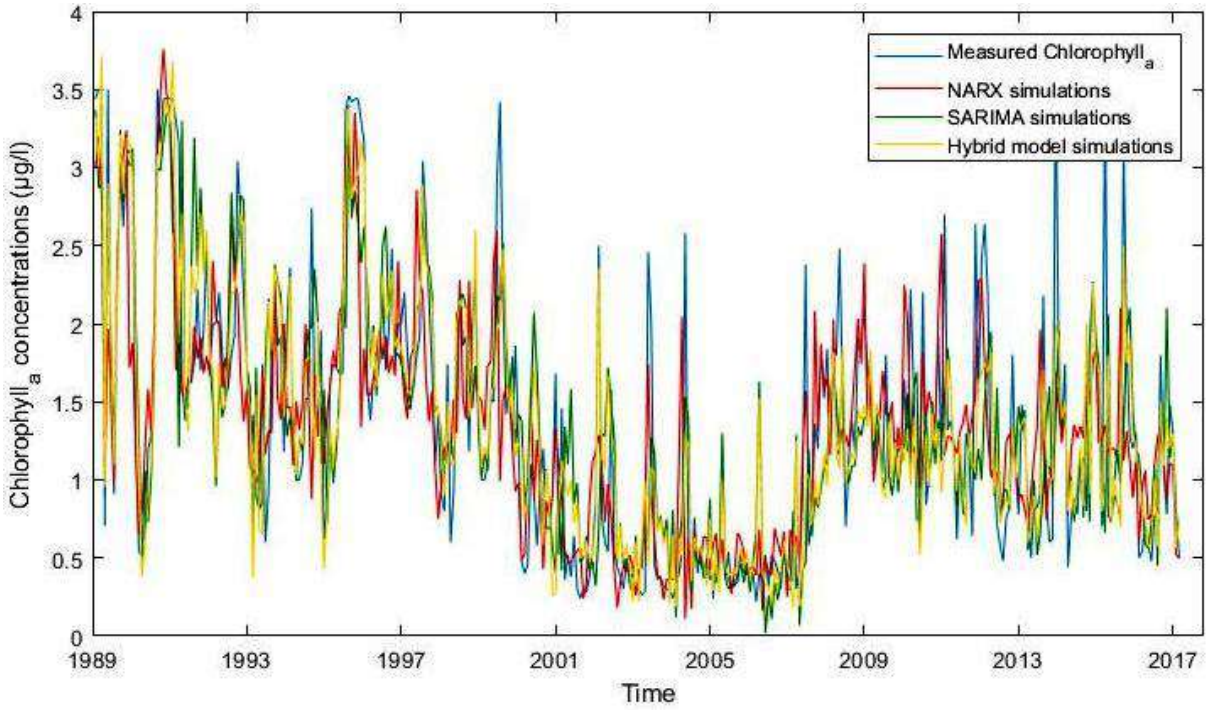



**Figure 68. Time series plot between observed and simulated values of Chlorophyll-*a* concentrations using SARIMA, NARX and hybrid approach in the North Lagoon of Tunis.**

Our findings demonstrated that the overall forecasting capability is improved when using the hybrid model. The comparison between the actual values and the forecast value using the hybrid model is given in figure 69.

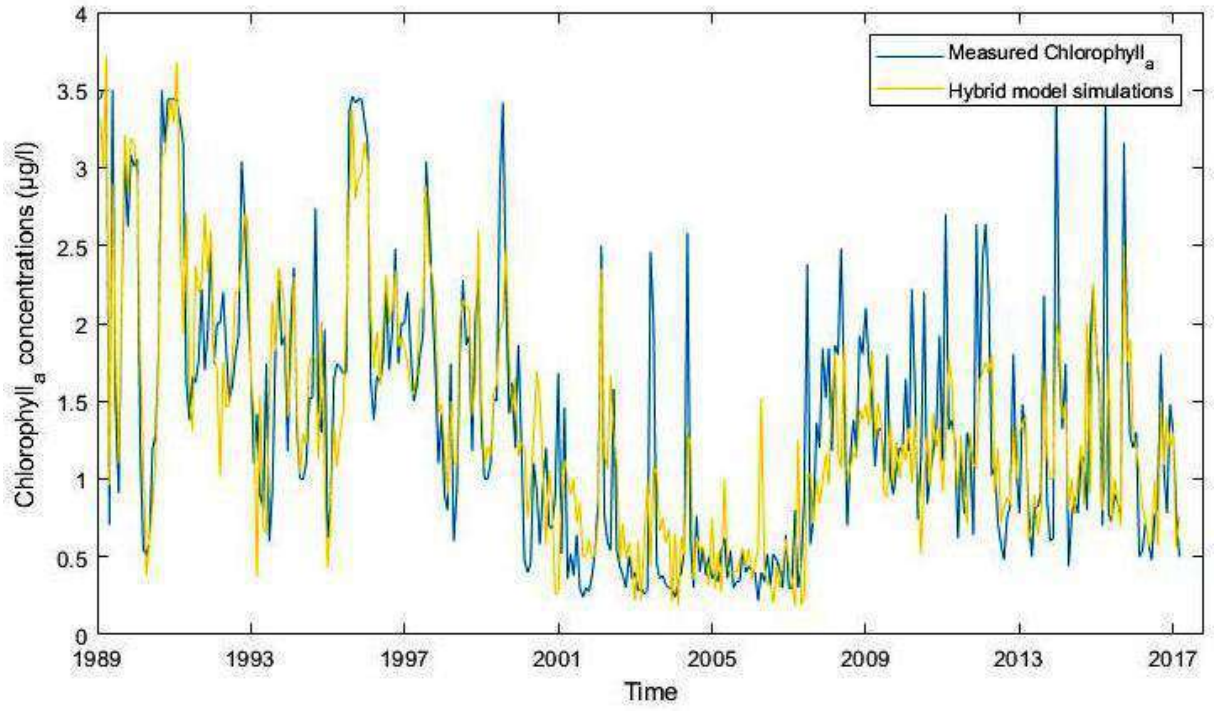



**Figure 69. Observed and modeled Chlorophyll-*a* variations using the hybrid model in the North Lagoon of Tunis.**

Results show that for short-term forecasting (1 month), both NARX network and hybrid models have much better accuracy than the SARIMA model. The hybrid model seems to predict relatively better the high concentrations values of the Chl-*a* than the NARX network, especially in approximately the five last years of the study period.

The forecasting of the Chl-*a* content one month ahead gave a value about 0.54 µg L-1, which was close to the observed value (0.50 µg L-1). The values are similar and reflect the relative usual Chl-*a* content of monthly variation in the lagoon, which means that this method is an effective one and can be used in the ecosystem under study.

## 6. Station 5

As in the previous chapters of this manuscript, for comparison, the hybrid approach was performed on Chl-*a* historical data of Station 5. Briefly, the results of the combined technique are presented in this section.

The predicting results for Chl-*a* concentrations in Station 5 using the various methods are shown in table 12.

**Table 14. Forecasting comparison for Chlorophyll-*a* concentrations in Station 5 using the different modeling approaches.**

| | MSE | R | $R^2$ |
|---|---|---|---|
| NARX network | 0.40 | 0.73 | 0.53 |
| SARIMA model | 0.44 | 0.66 | 0.45 |
| Hybrid model | 0.32 | 0.78 | 0.60 |

The results suggest that using NARX instead of the SARIMA model improves predicting accuracy. In comparison to the SARIMA model and the NARX network, the hybrid model's findings demonstrate that by combining two models, total forecasting errors may be considerably decreased.

Figure 70 gives the observed vs. simulated values with individual models of NARX and SARIMA as well as the hybrid technique. The hybrid model appears to predict high Chl-*a* concentrations better than the NARX network and SARIMA.

The results of the hybrid technique are satisfactory indicating that this approach is successful and applicable to the environment under investigation.

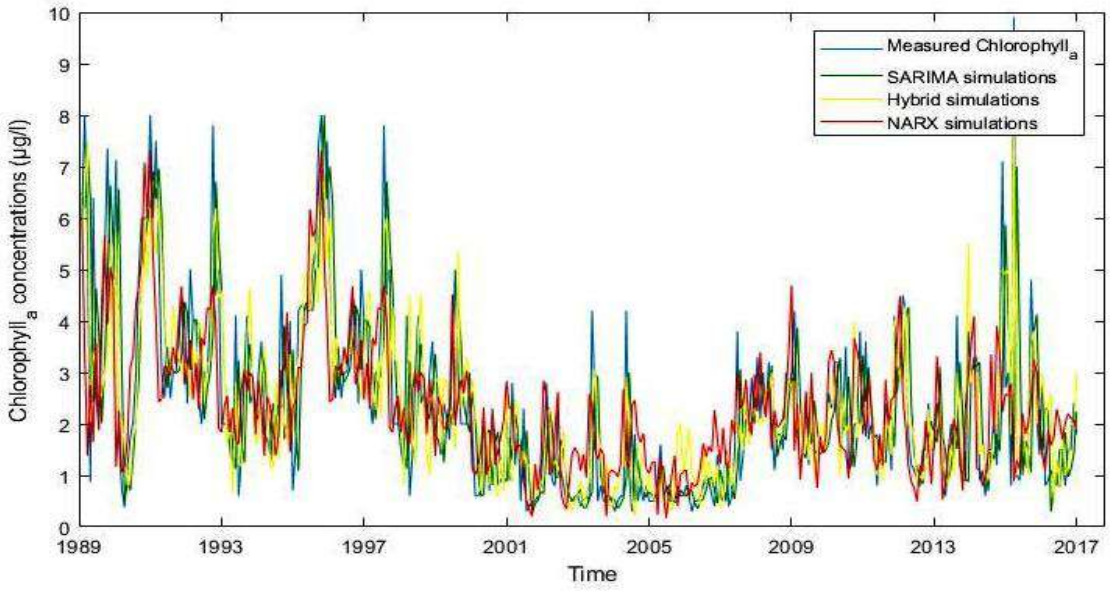



**Figure 70. Time series plot between actual and forecasted values of Chlorophyll-*a* concentrations in Station 5 using SARIMA, NARX and hybrid approach.**

# 7. Conclusion

Obtaining reliable predictions of a time series is an essential but challenging task. The accuracy of time series forecasting is critical for many decision processes, thus research to increase the efficiency of forecasting models has never stopped. The ARIMA model has become one of the most widely used forecasting techniques in both research and application. With their nonlinear modeling capabilities, artificial neural networks have lately demonstrated their potential in time series forecasting applications (Zhang, 2003). Despite the fact that both ARIMA and ANN are flexible in simulating a variety of problems, none is the optimal model for every forecasting case (Zhang, 2003).

As such, in this last chapter, we have proposed a hybrid forecasting method that applies SARIMA and NARX separately to model linear and nonlinear components, respectively of the Chl-*a* levels in the North Lagoon of Tunis and at Station 5 of the studied ecosystem. To the best of our knowledge, it is the first time that such approach is done in the mentioned lagoon.

Regarding the objective of the study to forecast the Chl-*a* concentrations with highest performance, various accuracy measurements have been tested. Comparing the three models using mathematic measurements (MSE, R and $R^2$), the hybrid model showed better skills in forecasting Chl-*a*, both at Station 5 and in the lagoon, than SARIMA model and NARX network.

SARIMA demonstrated some weakness in forecasting Chl-*a* fluctuations, especially for extreme values (minimum and maximum values) while the hybrid model has improved the forecasting accuracy in these points. The performances of the hybrid model and the NARX are relatively close with an improvement when using the combined method.

This combined approach could enable executives and managers in charge of the ecosystem to accurately estimate the monthly Chl-*a* concentrations as an eutrophication indicator in the

North Lagoon of Tunis. In addition, this integrated method might allow ecosystem executives and managers to properly predict monthly Chl-*a* concentrations in Station 5, the most affected area of the lagoon by the eutrophication.

This approach could be also applied to other natural ecosystems to estimate any key parameter, (such as nutrients) to evaluate the eutrophication level.

# Conclusion and Perspectives

**Problematic**

Recently, several factors have posed an increasing threat to the health of marine and coastal ecosystems. Natural and anthropogenic pressures, in addition to the effects of climate change, are posing serious risks to marine and coastal environments, affecting biological, chemical, and physical processes (Furlan *et al*., 2017).

Because Mediterranean coastal lagoon ecosystems are highly populated, they have been particularly vulnerable to human eutrophication in recent decades, owing mostly to urbanization (Zaldívar *et al*., 2008a; Souchu *et al*., 2010). Accordingly, ecological disturbances have been reported in different Mediterranean lagoons, estuaries and coastal water bodies, and also, all around the world (Sfriso *et al*., 2003; Heisler *et al*., 2008; Thornber *et al*., 2008). Among others, the North Lagoon of Tunis, subject of the present study. In fact, the North Lagoon of Tunis is one of the most important lagoons in Tunisia, which has experienced a critical ecological state mainly due to urban development.

Considering this, the ecological stability of the North Lagoon of Tunis makes it of significant socio-economic and ecological values. Multiples services are provided in this ecosystem, such as in tourism (water sports), in fisheries, and in the conservation of sea birds. Thus, it is necessary to improve our understanding of the eutrophication process and of the interactions among the water quality parameters in the lagoon, to adopt sustainable management strategies. In this study work, different forecasting approaches on the short and long term have been performed to evaluate the trophic state in the North Lagoon of Tunis considering the Chl-*a* as an eutrophication indicator.

In addition to the well-known Box and Jenkins approach, data-driven predictive models in the recent Machine Learning (ML) techniques were employed to predict Chl-*a* levels in the North

Lagoon of Tunis. To do so, a continuous monthly, approximately, three decades long, time series measurements (1989-2018) of seven physico-chemical parameters (water temperature, salinity, Secchi depth, dissolved oxygen, total phosphorus, total nitrogen and pH) and Chl-*a* were used. To the best of our knowledge, it is the first time that such approaches are done, especially, in the North Lagoon of Tunis.

**Defining underlying patterns in the time series in preparation for modeling work**

In the first chapter, a presentation of the study area and the spatio-temporal and seasonal variation of the parameters are performed.

In order to determine the most appropriate algorithm, the pre-processing of the data, as it was done in the first chapter, is an essential step to capture any underlying specificities such as spatial, temporal, seasonal, linear or nonlinear. Indeed, before doing any complex modeling, these specificities must be identified and taken into account.

Among the environmental and biological data, dissolved oxygen, total phosphorus and salinity concentrations with Chl-*a* values were most affected by seasonal variations.

In general, most parameters showed that the water quality in the North Lagoon was strongly influenced by the freshwater inflow from the Gulf of Tunis.

The results in this chapter confirmed the good functioning of the restoration project held in the North Lagoon of Tunis more than thirty years ago, based on the water circulation in the ecosystem. Indeed, it revealed that the coastal water was significantly influenced by freshwater input from the gulf of Tunis. However, given the fragility of this ecosystem, it remains highly influenced in particular at Station 5. The temporal variability is relatively similar among the Stations except for Station 5 (the furthest from the sea).

This environment must remain under surveillance and the biotic and abiotic components must be monitored regularly in the years to come. Knowing this, predictive modeling seems to be

an effective and realistic technique for monitoring the North Lagoon of Tunis and anticipating any potential deterioration.

It is important to mention that after studying the data in this chapter, it seemed heterogenic and nonlinear, which is expected in natural ecosystems. Considering this, using machine learning (ML) techniques seemed the most accurate methodologies to use. Indeed, ML has the most suitable approaches to handle this kind of data.

**Machine Learning based approaches**

The approach proposed in the second chapter, takes into account these specificities previously mentioned, and is based on a combination of ML methods, using firstly the random forest technique (RF) and secondly the nonlinear autoregressive with external inputs neural network algorithm (NARX) to predict and forecast on a short time scale the Chl-*a* concentrations in the North Lagoon of Tunis, as an indicator of eutrophication.

In order to be able to optimize the working time and reduce the cost, it is very important to limit the number of parameters to be measured. Therefore, in this study, it is of a big interest to select the specific variables that are most related to the Chl-*a* concentrations.

To do so, several input scenarios were tested, and the best one was derived from the Random Forest technique ($R^2$=0.62, MSE= 0.28), that showed that Secchi depth and dissolved oxygen are the two most important variables that condition the Chl-*a* variation in the studied ecosystem. Secchi depth and dissolved oxygen are appropriate variables that are very easy to measure without the need for additional laboratory analysis. These results can contribute in reducing the number of controlled parameters and thus save time and money.

Historical data of those two most important parameters, were used in addition to the Chl-*a* to predict one time step ahead (one month) the concentration of the eutrophication indicator, using another ML technique, which is the neural network with external inputs (NARX).

The developed NARX was able to predict the dynamics of Chl-*a* concentrations using a minimum of variables. The external inputs of NARX are the variables that contribute most to the variation of Chl-*a* according to the RF model (Secchi depth and dissolved oxygen).

Our results show that the eutrophication phenomenon could be modeled using the NARX technique and some extreme values were successfully estimated.

The developed NARX has an R= 0.79 and an MSE= 0.31. This proves that the predictions of our approach are sufficiently robust.

However, the results obtained from simulating the same approaches on the historical data of the Station 5 (the area the most affected by the eutrophication in the lagoon), presented a slight decrease in the performance of the models (R= 0.73). This can be due to the accentuated heterogeneity in this area of the lagoon. Nevertheless, RF and NARX models have great potential in modeling complex and heterogeneous systems, such as eutrophication in coastal lagoons.

This combined approach showed satisfactory performances. A good correlation was observed between the measured and predicted values. Machine Learning or Artificial Intelligence modeling is a promising and useful tool that optimizes monitoring techniques by identifying essential key variables and even Stations (thereby permitting cost and time reduction) and forecasts water quality variables with acceptable accuracy.

**Box and Jenkins Methodology**

The third chapter deals with the prediction of Chl-*a* as an indicator of eutrophication in the North Lagoon of Tunis, using the famous technique of Box and Jenkins (1967).

Long-term prediction model of Chl-*a* concentrations was developed using a seasonal autoregressive integrated moving average (SARIMA) method.

The SARIMA $(2,0,2)(2,0,2)_{12}$ model was the model with the best performance, with the lowest AIC and BIC, among a set of candidate ones.

The performance of the model was analyzed by comparing with actual data from our last observations and verifying the results of the residuals. The residuals were found to be uncorrelated and relatively normally distributed.

The forecast results are quite satisfactory (AIC= 628.91; BIC= 666.78; $R^2$=0.52) since the forecast period seems to reproduce relatively well the normal monthly Chl-*a* content in the lagoon.

For comparison, SARIMA technique was performed on Station 5 historical data. The results of the SARIMA $(1,1,0)(0,1,0)_{12}$ presented the best accuracy. However, The performance of the SARIMA $(1,1,0)(0,1,0)_{12}$ realized in Station 5 showed a small decline in comparison of the SARIMA $(2,0,2)(2,0,2)_{12}$ performed with the data of the lagoon. This might be owing to the lagoon's parameters of a big heterogeneity in this location.

**Hybrid technique**

In the previous chapters, SARIMA and NARX models individually, gave good results in predicting the linear and non-linear components of the Chl-*a* content in the North Lagoon of Tunis, respectively.

However, real problems generally have both linear and non-linear characteristics, and previous research has attempted to explain the variability of real problems by combining two methods.

In the last chapter, a novel hybrid linear-nonlinear methodology combining SARIMA and NARX network was implemented. The combined model led to an improvement and good performance in the result. The proposed hybrid model for the Chl-*a* parameter in the North Lagoon of Tunis showed an excellent coefficient of correlation R= 0.82.

The proposed hybrid model in this study has great potential to replace traditional models for predicting quality parameters and to boost monitoring operations.

The same combined technique was performed in historical data of Station 5 and presented the best forecasting results in comparison with NARX and SARIMA simulated in this area of the lagoon.

**The study contribution**

This study emphasizes the importance of predictive modeling in preventing and inhibiting eutrophication and therefore minimizing contamination in sensitive coastal ecosystem.

The North Lagoon of Tunis is presently a fully artificial environment as a result of human involvement, and ecological monitoring is required to ensure the proper ecological functioning of this ecosystem located in the heart of the metropolitan zone.

The capacity to automatically monitor water quality is extremely beneficial, particularly in sensitive regions where, there is a high risk of possible contamination incidents, and there are major socioeconomic activities that necessitate preventative intervention. To the best of our knowledge, no automated equipment exists that correctly measures Chl-*a* in real time.

Chl-*a* measurements must be performed in the laboratory which entails significant delay and expense (Jimeno-Sáez *et al*., 2020).

The modeling techniques described in this work for analyzing and forecasting the eutrophication problem in the North Lagoon of Tunis are especially beneficial for ecologists and environmentalists, as they will be able to anticipate water pollution levels and take required precautionary actions ahead of time. Also, these techniques can identify important parameters for enabling both selective physical/chemical monitoring and quick water quality assessment of the Tunis North Lagoon.

The developed models can be used to (1) estimate Chl-*a* concentrations when the real value is unavailable, (2) simulate alternative water quality scenarios over extreme ranges of input and output parameters, and (3) save money and time by eliminating sample cruises and laboratory testing wherever possible.

**Recommendations and perspectives**

Even though forecasting in a short time scale is very interesting in taking quick adequate decisions with more accurate results, it also represents the limitation of this study.

Further work on the algorithms, should be done to try to forecast on a long-time period in the future, with more accurate performances, using especially the hybrid technique.

As recommendations for further research, many scenarios can be performed using these models. From the most optimistic to the most pessimistic, by implementing the adequate values of the Chl-*a* or any other key parameter that can describe eutrophication process in natural ecosystems.

In order to improve the accuracy of the models we would suggest, adding more data either by simulation (in interpolating the available data), or ideally by performing daily or weekly measurements, at least for the most important parameters (Chl-*a*, dissolved oxygen, nutrients), maybe not at all the Stations but at a minimum at the two Stations 2 (in the north area of the lagoon) and 5 (at the south area of the lagoon and the furthest from the sea).

Station 4 is the shallowest area of the lagoon. For this reason, it is also preferable to continue monitoring in Station 4 in addition to Station 5. The exchanges between the atmosphere and the water are very enhanced due to the low depth of the ecosystem in Station 4. For instance, high temperatures usually lead to a decrease in the dissolved oxygen level in natural ecosystems.

The findings verified the significance and use of intelligent ML modeling as a quick, simple, and cost-effective technique.

Finally, it is worth noting that these types of techniques are often utilized to forecast the progression of the world's largest crisis such as the Coronavirus pandemic.

# Bibliographical references


Abba S.I, Hadi S.J, Abdullahi J (2017) River water modeling prediction using multi-linear regression, artificial neural network and adaptive neuro-fuzzy inference system techniques. Procedia Computer Science 120: 75-82. https://doi.org/10.1016/j.procs.2017.11.212.

Adhikari R, Agrawal RK (2013) An introductory study on time series modeling and forecasting. LAP Lambert Academic Publishing, Germany. 67 pages. arXiv:1302.6613.

Afli A, Ayari R, Brahim M (2008) Trophic organization of the macro-zoobenthic assemblages within coastal areas subjected to anthropogenic activities. J. Mar. Biol. Assoc. UK 88:663-674. https://doi.org/10.1017/S0025315408001318.

Agrawal A (2011) A new approach to spatio temporal kriging and its application. Thesis master of science. Graduate School of the Ohio State University, p 112.

Aho K, Derryberry D, Peterson T (2014) Model selection for ecologists: the worldviews of AIC and BIC. Ecology 95(3):631-636. https://doi.org/10.1890/13-1452.1.

Akaike H (1972) Use of an information theoretic quantity for statistical model identification. In Proceedings of the 5th Hawaii International Conference on System Sciences, pp 249–250.

Albertson K, Aylen J (1996) Modelling the great lakes freeze: forecasting and seasonality in the market for ferrous scrap. Int J Forecast 12(3):345–359. https://doi.org/10.1016/0169-2070(96)00669-3.

Allard R (1998) Use of time-series analysis in infectious disease surveillance. Bull World Health Organ 1998 76(4):327-333.

Al Shehhi, MR., Kaya, A., 2020. Time series and machine learning to forecast the water quality from satellite data. 21 pages. arXiv:2003.11923.

Aminot A, Chausseppied M (1983) Manuel des analyses chimiques en milieu marin. Centre National pour l'éxploitation des Océans, 395 p.

APHA (1999) Standard Methods for the Examination of Water and Wastewater. American Public Health Association, Washington DC. 1268 p.

Armi Z, Trabelsi E, Turki S, Ben Maïz N, Mahmoudi E (2012) Composition and dynamics of potentially toxic dinoflagellates in a shallow Mediterranean lagoon. Oceanol. Hydrobiol. Stud. 41(3): 25–35.

Armstrong, J.S (1985) Long range forecasting: from crystal ball to computer. 2nd edition. John Wiley, New York. pp 689.

ASCE Task Committee (2000) Artificial neural networks in hydrology I. Preliminary concepts. Journal of Hydrology and Engineering 5 :115-123. https://doi.org/10.1061/(asce) 1084-0699(2000)5:2(115).

Babovic V, Sannasiraj SA, Soon Chan E (2005) Error correction of a predictive ocean wave model using local model approximation. Journal of Marine Systems; 53:1-17. https://doi.org/10.1016/j.jmarsys.2004.05.028.

Barnes RSK (1980) Coastal Lagoons: The Natural History of a Neglected Habitat; Cambridge University Press: Cambridge, UK. 106 pp.

Béjaoui B, Armi Z, Ottaviani ., Barelli E, Gargouri-Ellouz E, Chérif R, Turki S, Solidoro C, Aleya L (2016) Random Forest model and TRIX used in combination to assess and diagnose the trophic status of Bizerte Lagoon, southern Mediterranean. Ecological Indicators 7: 293-301. http://dx.doi.org/10.1016/j.ecolind.2016.07.010.

Béjaoui B, Ottaviani E, Barelli E, Ziadi B, Dhib A, Lavoie M, Gianluca C, Turki S, Solidoro C, Aleya L (2018) Machine learning predictions of trophic status indicators and plankton dynamic in coastal lagoons. Ecological Indicators 95:765-774. https://doi.org/10.1016/j.ecolind.2018.08.041.

Belkhir M (1980) Eutrophisation du lac Nord de Tunis : étude physico-chimiques et biomasses phytoplanctonique et macroalgue. Thèse 3° cycle, Fac. Sci. Tunis, p 194.

Belkhir M, Hadj Ali Salem M (1981) Contribution à l'étude des mécanismes d'eutrophisation dans le lac de Tunis: évolution des paramètres physico-chimiques et biologiques. Bull. Inst. Nation. Scient. Tech. Océanogr. Pêche Salammbo, 10 :5-26.

Belkhir M (1984) Dynamique des peuplements algaux dans le lac de Tunis. Bull. Inst. Nat. Scient. Tech. Océanogr. Pêche Salammbô 11:41-69.

Ben Charrada R (1992) Le lac de Tunis après les aménagements. Paramètres physicochimiques de l'eau et relation avec la croissance des macroalgues Marine Life 1: 29-44.

Bennett M.G, Schofield K.A, Lee S.S, Norton S.B (2017) Response of chlorophyll *a* to total nitrogen and total phosphorus concentrations in lotic ecosystems: a systematic review protocol. *Environ Evid (2017) 6:18.* DOI 10.1186/s13750-017-0097-8.

Ben Maiz N (1992) Analyse de la qualité des eaux et évolution des peuplements végétaux du lac nord de Tunis. Rapport interne. Société de Promotion du lac de Tunis, pp. 22 Annexes 1 et 2.

Ben Maiz N (1995) Etude nationale sur la biodiversité biologique de la flore marine et aquatique en Tunisie (Monographie). Projet de coopération MEAT/PNUE/GEF, Ministère de l'environnement, Tunisie, 78 pp.

Ben Maiz N (1997) Le Lac Nord de Tunis : un milieu en mutation. In Gestion et conservation des zones humides tunisiennes. Actes de séminaire. 77-84.

Ben Maiz N (2008) Le milieu naturel du Lac Nord de Tunis. In Le Lac Nord de Tunis et le fort « Chikly » - De l'eau jaillit la vie et se développent les civilisations. Edition SPLT 2008, Imprimerie Simpact, pp 96-117.

Benihabib ME, Ahmadian A, Jamali FS (2017) Hybrid DARIMA-NARX model for forecasting long-term daily inflow to Dez reservoir using the North Atlantic Oscillation (NAO) and rainfall data. GeoResJ 13:9-16. https://doi.org/10.1016/j.grj.2016.12.002.

Borja A, Dauer DM (2008) Assessing the environmental quality status in estuarine and coastal systems: comparing methodologies and indices. Ecol Indicat 8(4):331–337. https://doi.org/10.1016/j.ecolind.2007.05.004.

Bowden GJ, Nixon JB, Dandy GC, Maier, HR, Holmes M (2006) Forecasting chlorine residuals in a water distribution system using a general regression neural network. Mathematical and Computer Modelling 44:469–484. https://doi.org/10.1016/j.mcm.2006.01.006.

Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control. 3rd ed. Holden-Day, California. 553 p.

Box GEP, Jenkins GM, (1976) Time Series Analysis: Forecasting and Control. Holden-Day, San Francisco, CA. 575 p.

Box GEP, Jenkins GM, Reinsel GC, Ljung GM, (2008) Time Series Analysis: Forecasting and Control, 4th edition, John Wiley & Sons, Hoboken, NJ, USA. 755 p.

Breiman L (2001) Random Forests. Machine Learning 45: 5–32. https://doi.org/10.1023/A:1010 10 933404324.

Breitburg D (2002) Effects of hypoxia, and the balance between hypoxia and enrichment, on coastal fishes and fisheries. Estuaries 25:767–781. https://doi.org/10.1007/ BF02804904.

Brown PC (1992) Spatial and seasonal variation in chlorophyll distribution in the upper 30 m of the photic zone in the southern Benguela/Agulhas ecosystem. South African Journal of Marine Science 12(1):515-525. https://doi.org/10.2989/02577619209504722.

Chapman PM, Paine MD, Arthur AD, Taylor LA (1996) A triad study of sediment quality associated with a major, relatively untreated marine sewage discharge. Mar. Pollut. Bull. (32): 47–64.

Charulatha G, Srinivasalu S, Uma Maheswari O *et al*. (2017) Evaluation of ground water quality contaminants using linear regression and artificial neural network models. Arab J Geosci 10:128. https://doi.org/10.1007/s12517-017-2867-6.

Chatfield C (1988) What is the ‘best’ method of forecasting? J. Appl. Statist. (15):19–39.

Chen KY, Wang CH (2007) A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. Expert Syst. Appl. 32 (1): 254–264.

Chen Q, Guan T, Yun L, Li R, Recknagel F (2015) Online forecasting chlorophyll a concentrations by an auto-regressive integrated moving average model: Feasibilities and potentials. Harmful Algae 43:58–65. http://dx.doi.org/10.1016/j.hal.2015.01.002.

Clemen R (1989) Combining forecasts: a review and annotated bibliography with discussion, Int. J. Forecasting 5:559–608.

Cloern JE (2001) Our evolving conceptual model of the coastal eutrophication problem. Mar. Ecol. Prog. Ser. 210:223-253. https://doi.org/10.3354/meps210223.

Codling EA, Dumbrell AJ (2012) Mathematical and theoretical ecology: linking models with ecological processes. The Royal Society Interface Focus. 2(2): 144–149 https://doi.org/10.1098/rsfs.2012.0008.

Coelho HF, Canter PH, Ernst E (2007) Mindfulness-based cognitive therapy: evaluating current evidence and informing future research. J Consult Clin Psychol. 75(6):1000-1005. https://doi.org/10.1037/0022-006X.75.6.1000. PMID: 18085916.

Cowpertwait SPP, Metcalfe VA (2008) Introductory to time series with R .Springer Dordrecht Heidelberg London, New York. p 249.

Cui L, You L, Huang Z (2007) TN/TP ratio of lake water and its implication in algae bloom of Beijing’s urban lakes. Environ. Sci. Technol. 30(10):47–49.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random forests for classification in ecology. Ecology 88: 2783-2792. https://doi.org/10.1890/07-0539.1.

Daliakopoulos IN, Coulibaly P, Tsanis IK (2005) Groundwater level forecasting using artificial neural networks. Journal of Hydrology 309: 229-240. https://doi.org/10.1016/j. jhydrol.2004.12.001.

David M, Bailly-Comte V, Munaron D, Fiandrino A, Stieglitz TC (2019) Groundwater discharge to coastal streams – A significant pathway for nitrogen inputs to a hypertrophic Mediterranean coastal lagoon. Science of the Total Environment 677:142–155. https://doi.org/10.1016/j.scitotenv.2019.04.233.

Davis TW, Berry DL, Boyer GL, Gobler CJ (2009) The effects of temperature and nutrients on the growth and dynamics of toxic and non-toxic strains of Micro-cystis during cyanobacteria blooms. Harmful Algae 8:715–725. https://doi.org/10.1016/j .hal.2009.02.004.

de Jonge VN, Elliott M, Orive E (2002) Causes, historical development, effects and future challenges of a common environmental problem: eutrophication. Hydrobiologia 475/476:1–19. https://doi.org/10.1023/A:1020366418295.

Delbaere B, Nieto-Serradilla A (2004) Environmental Risks from Agriculture in Europe: Locating Environmental Risk Zones in Europe Using Agri-Environmental Indicators. – Tilburg, ECNC-European Centre for Nature Conservation. p 184 p.https://lib.ugent.be/catalog/rug01:000933733.

Derolez V, Soudant D, Malet N, Chiantella C, Richard M, Abadie E, Aliaume C, Bec B (2020) Two decades of oligotrophication: Evidence for a phytoplankton community shift in the coastal lagoon of Thau (Mediterranean Sea, France). Estuarine, Coastal and Shelf Science 241: 106 810. https://doi.org/10.1016/j.ecss.2020.106810.

Desmit X, Ruddick K, Lacroix G (2015) Salinity predicts the distribution of chlorophyll-a spring peack in the southern North Sea continental waters. Journal of Sea Research 103 (2015) 59-74. https://doi.org/10.1016/j.seares.2015.02.007.

Dindarloo S (2015) Reliability forecasting of a load-haul-dump machine: a comparative study of ARIMA and neural networks. Qual. Reliab. Eng. Int. 32(4):1545–1552. https://doi.org/10.1002/qre.1844.

Dippner JW, Nguyen-Ngoc L, Doan-Nhu H, Subramaniam A (2011) A model for the prediction of harmful algae bloom in the Vietnamese upwelling area. Harmful Algae 10:606–611. https://doi.org/10.1016/j.hal.2011.04.012.

Dogan E, Sengorur B, Koklu R (2009) Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. J. Environ. Manag. 90:1229–1235.

Domingos P, Gômara GA, Sampaio GF, Soares MF, de Freitas Lopes Soares F (2012) Fish mortality events associated with phytoplanktonic flowers in the Rodrigo De Freitas Lagoon: 10-year monitoring program. Oecologia Australis 16(03):441-466. DOI: 10.4257/oeco.2012.1603.09.

Donner A (1982) The relative effectiveness of procedures commonly used in multiple regression analysis for dealing with missing values. Am. Stat. 36:378–381.

Dopazo J, Huaichun W, Carazo JM (1997) A new type of unsupervised growing neural network for biological sequence classification that adopts the topology of a phylogenetic tree. Lect. Notes Comput. Sci. 1240:932–941.

Drius M, Bongiorni L, Depellegrin D, Menegon S, Pugnetti A, Stifter S (2018) Tackling challenges for Mediterranean sustainable coastal tourism: An ecosystem service perspective. Science of the Total Environment 652:1302–1317. https://doi.org/10.1016/j.scitotenv.2018.10.121.

Elliott M, Quintino V (2007) The Estuarine Quality Paradox, Environmental Homeostasis and the difficulty of detecting anthropogenic stress in naturally stressed areas. Marine Pollution Bulletin 54(6):640-5. https://doi.org/10.1016/j.marpolbul.2007.02.003.

Ferreira JG, Andersen JH, Borja A et al. (2011) Overview of eutrophication indicators to assess environmental status within the European Marine Strategy Framework Directive. Estuarine, Coastal and Shelf Science 93 : 117-131.

Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. 12:42–58. https://doi.org/10.2307/1411 (doi:10.2307/1411).

Fogelman S, Zhao H, Blumenstein M, Zhang S (2006) Estimation of oxygen demand levels using UV-Vis spectroscopy and artificial neural networks as an effective tool for real-time, wastewater treatment control. In Proceedings of the 1st Australian Young Water Professionals Conference, Sydney, Australia.

Fraley C, Raftery AE (1998) How many clusters? Which clustering method? Answers via model-based cluster analysis. Comput. J. 41:578–588.

Frolov S, Rayan JP, Chavez FP (2012) Predicting eutrophic-depth-integrated chlorophyll-a from discrete-depth and satellite-observable chlorophyll-a off central California. J. Geophys. Res., 117, C05042, doi:10.1029/2011JC007322.

Furlan E, Torresan S, Critto A, Marcomini A (2017) Spatially explicit risk approach for multi-hazard assessment and management in marine environment: The case study of the Adriatic Sea. Science of the Total Environment 618:1008–1023. https://doi.org/10.1016/j.scitotenv.2017.09.076.

García-Ayllón S (2017) Diagnosis of complex coastal ecological systems: Environmental GIS analysis of a highly stressed Mediterranean lagoon through spatiotemporal indicators. Ecological Indicators. 83:451-462. http://dx.doi.org/10.1016/j.ecolind.2017.08.015.

García-Pintado J, Martínez-Mena M, Barberá GG, Albaladejo J, Castillo VM (2007) Anthropogenic nutrient sources and loads from a Mediterranean catchment into a coastal lagoon: Mar Menor, Spain. Science of the Total Environment 373:220–239. https://doi.org/10.1016/j.scitotenv.2006.10.046.

Garson GD (1991) Interpreting neural network connection weights. Artif. Intell. Expert 6:47–51.

Gaynor PE, Kirkpatrick RC (1994) Introduction to Time Series Modeling and Forecasting in Business and Economics. McGraw-Hill college. p 625.

Gershenfeld NA, Weigend AS (1993) The Future of Time Series. In: Weigend AS Gershenfeld NA (ed) Time Series Prediction: Forecasting the Future and Understanding the Past, eds. Addison, Wesley, pp 1–70.

Ginzburg I, Horn D (1994) Combined neural networks for time series analysis, Adv. Neural Inf. Process. Systems 6:224–231.

Griitzner R (1996) Environmental modeling and simulation - applications and future requirements. Computer Science. Corpus ID: 59668493.

Gohel MS, Heatly F, Liu X et al. (2019) A randomized trial of early endovenous ablation in venous ulceration. N Engl J Med 378:2105-114. https://doi.org/10.1056/NEJM oa1801214.

Haizum Abd Rahman N, Hisyam Lee M, Suhartono S, Latif, MT (2019) Hybrid Seasonal ARIMA and Artificial Neural Network in Forecasting Southeast Asia City Air Pollutant Index. ASM Science Journal 12(1):215-226.

Håkanson L, Eklund Jenny M (2010) Relationships Between Chlorophyll, Salinity, Phosphorus, and Nitrogen in Lakes and Marine Areas. Journal of Coastal Research 26(3):412-423. DOI:10.2112/08-1121.1.

Harbridge W, Pilkey OH, Whaling P, Swetland P (1976) Sedimentation in the lake of Tunis: a lagoon strongly influenced by man. Environmental Geology 1:215-225. https://doi. org/10.1007/bf02407508.

Hasting A (2001) Transient dynamics and persistence of ecological systems. Ecol. Lett. 4:215–220. https://doi.org/10.1046/j.1461-0248.2001.00220.x.

Hassanzadeh S, Hosseinibalam F, Alizadeh R (2009) Statistical models and time series forecasting of sulfur dioxide: a case study Tehran. Environ Monit Assess 155(1):149–155. https://doi. org/10.1007/s10661-008-0424-1.

Hayajneh MT, Hassan AM, Mayyas AT (2009) Artificial neural network modeling of the drilling process of self-lubricated aluminum/alumina/graphite hybrid composites synthesized by powder metallurgy technique. Journal of Alloys and Compounds 478:559–565.

He Z, Wen X, Liu H, Du J (2014) A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. Journal of Hydrology 509: 379-386. https://doi.org/ 10.1016/j.jhydrol.2013.11.054.

Heisler J, Glibert PM, Burkholder JM et al. (2008) Eutrophication and harmful algal blooms: A scientific consensus. Harmful Algae 8:3–13. https://doi.org/ 10.1016/j.hal .2008.08.006.

Helfenstein U (1991) The use of transfer function models, intervention analysis and related time series methods in epidemiology. Int J Epidemiol 20(3): 808-815. doi: 10.1093/ije/20.3.808.

Herrera-Silveira Jorge A, Medina-Gomez I, Colli R (2002) Trophic status based on nutrient concentration scales and primary producers community of tropical coastal lagoons influenced by groundwater discharges. Hydrobiologia 475: 91–98.

Hill DJ, Minsker BS (2010) Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ. Model Softw. 25:1014–1022.

Hintze J (2007). NCSS 2007. NCSS, LLC. Kaysville, Utah, USA. pp 585.

Hong T, Koo C, Park S (2012) A decision support model for improving a multi-family housing complex based on CO2 emission from gas energy consumption. Build Environ 52:142-51.

Hood RR, Zhang X, Glibert PM, Roman MR, Stocker DK (2006) Modeling the influence of nutrients, turbulence and grazing on Pfiesteria population dynamics. Harmful Algae 5:459–479. https://doi.org/10.1016/j.hal.2006.04.014.

Hornik K, Stinchocombe M, White H (1989) Multilayer feedforward networks are universal approximators, Neural Networks 2:359–366.

Howarth RW, Marino R (2006) Nitrogen as the Limiting Nutrient for Eutrophication in Coastal Marine Ecosystems: Evolving Views over Three Decades. Limnology and Oceanography 51(1_part_2):364-376. https://doi.org/10.4319/lo.2006.51.1_part_2.0364.

Huang J, Gao J, Zhang Y (2015) Combination of artificial neural network and clustering techniques for predicting phytoplankton biomass of lake Poyang, China. Limnology 16 :179-191. https://doi.org/ 10.1007/s10201-015-0454-7.

Hwang HB (2001) Insights into neural-network forecasting time series corresponding to ARMA (p; q) structures, Omega 29: 273–289.

Hyndman RJ, Athanasopoulos G (2013) Forecasting: Principles and Practice. O Texts Publishers. pp 138. Available online at: https://www.otexts.org/book/fpp.

IOC, SCOR, IAPSO (2010) The international thermodynamic equation of seawater – 2010: Calculation and use of thermodynamic properties. Intergovernmental Oceanographic Commission, Manuals and Guides No. 56, UNESCO (English), 196 pp.

Jenkins GM (1982) Some practical aspects of forecasting in organisations, J. Forecasting 1: 3–21.

Jeong K, Koo C, Hong T (2014) An estimation model for determining the annual energy cost budget in educational facilities using SARIMA (seasonal autoregressive integrated moving average) and ANN (artificial neural network). Energy 71:71-79. http://dx.doi.org/10.1016/j.energy.2014.04.027.

Jimeno-Sáez P, Senent-Aparicio, JM, Cecilia J, Pérez-Sánchez J (2020) Using Machine-learning algorithms for eutrophication modeling: case study of Mar Menor lagoon (Spain). International Journal of Environmental Reseach and Public Health 17: 1189. https://doi.org/ 10.3390/ijerph17041189.

Jouirou M (1982) Faciès sédimentaires et processus dynamiques dans la formation d'un milieu lagunaire : évolution Holocène et actuelle du lac de Tunis et ses abords. Thèse de Doctorat en géologie et applications au domaine marin, Université de Bordeaux 1, 218 p.

Juntunen P, Liukkonen M, Pelu M, Lehtola M, Hiltunen Y (2012) Modelling of water Quality: an application to a water treatment process. Applied Computational Intelligence and Soft Computing 9 pages. https://doi.org/10.1155/2012/846321.

Kalaji HM, Sytar O, Brestic M, Samborska IA, Cetner MD, Carpentier C (2016) Risk assessment of urban lake water quality based on in situ cyanobacterial and total chlorophyll-a monitoring. Pol. J. Environ. Stud 25(2): 655–661. https://doi.org/10.15244/pjoes/60895.

Khandelwal I, Adhikari R, Verma G (2015) Time Series Forecasting using Hybrid ARIMA and ANN Models based on DWT Decomposition. Procedia Computer Science 48: 173 – 179.

Kinné O (1958) Adaptation to salinity variations - some facts and problems. In: Physiological Adaptation; edited by C. L. Prosser. Am. Physiol. Soc. Washington. D. C., 92-106.

Kohavi R, John GH (1997). Wrappers for feature subset selection. Artificial Intelligence, 97 (1–2):273–324.

Kompare B, Bratko I, Steinman F, D~eroski S (1994) Using machine learning techniques in the construction of models. Ecological Modelling 75/76:617-628.

Koo Choong- Wan, Hong TaeHoon, Hyun Chang- Taek et al. (2010) A study on the development of a cost model based on the owner's decision making at the early stages of a construction project. International Journal of Strategic Property Management, 14(2):121-137. https://doi.org/ 10.3846/ijspm.2010.10.

Kuo JT, Hsieh MH, Lung WS, She N (2007) Using Artificial Neural Network for reservoir eutrophication prediction. Ecol. Model. 200:171–177.

Latif Z, Tasneem MA, Javed T, Butt S, Fazil M, Ali M, Sajjad MI (2002) Evaluation of water quality by Chlorophyll and Dissolved Oxygen. Science Vision, 7(3-4), 160-168.

Lee JHW, Huang Y, Dickmen M, Jayawardena AW (2003) Neural network modelling of coastal algal blooms. Ecological Modelling 159: 179-201. https://doi.org/10.1016 /S0304-3800(02)00281-8.

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

Lek S, Guégan JF (1999) Artificial neural networks as a tool in ecological modelling, an introduction. Ecological modelling 120:65–73.

Leruste A, Malet N, Munaron D, Derolez V, Hatey E, Collos Y, De Wit R, Bec B (2016) First steps of ecological restoration in Mediterranean lagoons: Shifts in phytoplankton communities. Estuarine, Coastal and Shelf Science 180:190-203. http://dx.doi.org/10. 1016/j.ecss.2016.06.029.

Levin SA (2001) Encyclopedia of Biodiversity. Academic Press, San Diego, CA. Volume I. 943 pp.

Li X, Sha J, Wang ZL (2017) Chlorophyll-A Prediction of Lakes with Different Water Quality Patterns in China Based on Hybrid Neural Networks. Water 9:524.

Lin T, Horne BG, Tino P, Giles CL (1996) Learning long-term dependencies in NARX recurrent neural networks. IEEE Transactions on Neural Networks 7(6): 1329 - 1338. https://doi.org/10.1109/72.548162.

Linthicum KJ, Anyamba A, Tucker CJ, Kelley PW, Myers MF, Peters CJ (1999) Climate and satellite indicators to forecast Rift Valley fever epidemics in Kenya. Science 285(5426): 397-400. https://doi.org/10.1126/science.285.5426.397.

Lloret J, Marin A, Marin-Guirao L (2008) Is coastal lagoon eutrophication likely to be aggravated by global climate change? Estuarine, Coastal and Shelf Science 78:403-412. https://doi.org/10.1016/j.ecss.2008.01.003.

Lorenzen CJ (1967) Determination of chlorophyll and pheopigments by spectrophotometric equations. Limnol. Oceanogr. 12: 34–346.

Lu Y, AbouRizk SM (2008) Automated Box–Jenkins forecasting modelling. Automation in Construction 18: 547–558.

Lu F, Chen Z, Liu W, Shao H (2016) Modeling chlorophyll-a concentrations using an artificial neural network for precisely eco-restoring lake basin. Ecological Engineering 95: 422-429. http://dx.doi.org/10.1016/j.ecoleng.2016.06.072.

Lui GCS, Li WK, Leung KMY, Lee JHW, Jayawardena AW (2007) Modelling algal blooms using vector autoregressive model with exogenous variables and long memory filter. Ecol. Model. 200:130–138. https://doi.org/10.1016/j.ecolmodel.2006.06.017.

MacArthur RH, Wilson EO (1967) The theory of island biogeography. Princeton, NJ: Princeton University Press. p 181.

Maier HR, Jain A, Dandy GC, Sudheer K (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environ. Model. Softw 25: 891–909. https://doi.org/10.1016/j .envsoft.2010.02.003.

Makridakis S, Anderson A, Carbone R, Fildes R, Hibdon M, Lewandowski R, Newton J, Parzen E, Winkler R (1982) The accuracy of extrapolation (time series) methods: results of a forecasting competition, J. Forecasting 1:111–153.

Makridakis S, Chatfield C, Hibon M, Lawrence M, Millers T, Ord K, Simmons LF (1993) The M-2 competition: a real-life judgmentally based forecasting study, Int. J. Forecasting 9:5–29.

Makwinja R, Phiri T, Kosamu IB, Kaonga CC (2017) Application of stochastic models in predicting Lake Malawi water levels. Int J Water Resour Environ Eng 9(9):191–200. DOI: 10.5897/IJWREE2017.0740.

Mammadli S (2017) Financial time series prediction using artificial neural network based on Levenberg-Marquardt algorithm. Procedia Comput. Sci. 120:602–607. https:// doi.org/10.1016/j.procs.2017.11.285.

Marine Strategy Framework Directive (2008) Directive 2008/56/EC of the European Parliament and of the Council of 17 June 2008 establishing a framework for community action in the field of marine environmental policy (Marine Strategy Framework Directive) (Text with EEA relevance). Official Journal of the European Union EUR-Lex - 32008L0056 - EN L 164:19.

Markova M (2019) Foreign Exchange Rate Forecasting by Artificial Neural Networks. In AIP Conference Proceedings 2164:060010. pp 1-14.

Mathier L, Fagherazzi L, Rassam JC, Bobée B (1992) Great lakes net basin supply simulation by a stochastic approach. Rapport No. 362, INRS-Eau. Université du Québec. pp 104.

Mc Quaid N, Zamyadi A, Prevost M, Bird DF, Dorner S (2011) Use of in vivo phycocyanin fluorescence to monitor potential microcystin-producing cyanobacterial biovolume in a drinking water source. Journal of Environmental Monitoring 13 :455-463. https://doi.org/10.1039/c0em00163e. Epub 2010 Dec 15.

Mdaini Z, El Gafsi M, Tremblay J, Pharand P, Gagné JP (2019) Spatio-temporal variability of biomarker responses and lipid composition of Marphysasanguinea, Montagu (1813) in the anthropic impacted lagoon of Tunis. Marine Pollution Bulletin 144:275-286. https://doi.org/10.1016/j.marpolbul.2019.04.065.

Menendez M, Martınez M, Comın FA (2001) A comparative study of the effect of pH and inorganic carbon resources on the photosynthesis of three floating macroalgae

species of a Mediterranean coastal lagoon. Journal of Experimental Marine Biology and Ecology 256:123–136.

Mishra AK, Desai VR (2005) Drought forecasting using stochastic models. Stoch. Environ. Res. Risk Assess. 19:326–339.

Mitchell MW (2011) Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters. Open Journal of Statistics 1: 205-211. https://doi.org/10.42 36/ojs.2011.13024.

Mjalli FS, Al-Asheh S, Alfadala HE (2006) Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance. J. Environ. Manag 83:329–338. https://doi.org/10.1016/j.jenvman.2006.03.004.

Mooney H, Larigauderie A, Cesario M, Elmquist T, Hoegh-Guldberg O, Lavorel S, Mace GM, Palmer M, Scholes R, Yahara T (2009) Biodiversity, climate change, and ecosystem services. Current Opinions in Environmental Sustainability 1:46-54. https://doi.org/10.1016/J.COSUST.2009.07.006.

Motoda H, Liu H (2002) Feature selection, extraction and construction. Towards the Foundation of Data MiningWorkshop. In Proceedings of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'02), Taipei, Taiwan; pp. 67–72.

Mulia IE, Tay H, Roopsekhar K, Tkalich P (2013) Hybrid ANN–GA model for predicting turbidity and chlorophyll a concentrations. Journal of Hydro-Environmental Research 7: 279-299. https://doi.org/10.1016/j.jher.2013.04.003.

Nayak PC, Sudheer KP, Rangan DM, Ramasastri KS (2005) Short-term flood forecasting with a neuro fuzzy model. Water Resources Research 41:2517-2530. https://doi.org/10.1 029/2004WR003562.

Nazeer M, Wong MS, Nichol JE (2017) A new approach for the estimation of phytoplankton cell counts associated with algal blooms. Science of the Total Environment 590-591:125-183. http://dx.doi.org/10.1016/j.scitotenv.2017.02.182.

Newbold P, Granger CWJ (1974) Experience with forecasting univariate time series and the combination of forecasts (with discussion), J. R. Statist. Soc. Ser. A 137: 131–164.

Newton A, Icely JD, Falcao M, Nobre A, Nunes JP, Ferreira JG, Vale C (2003) Evaluation of eutrophication in the Ria Formosa coastal lagoon, Portugal. Continental Shelf Research 23:1945–1961.

Newton A, Icely J, Cristina S et al. (2014) An overview of ecological status, vulnerability and future perspectives of European large shallow, semi-enclosed coastal systems, lagoons and transitional waters. Estuar. Coast. Shelf Sci 140: 95–122. https://doi.org/10.1016/J.ECSS.2013.05.023.

Newton A, Brito AC, Icely JD et al (2018) Assessing, quantifying and valuing the ecosystem services of coastal lagoons. Journal of Nature Conservation 44: 50-56. https://doi.org/10.1016/j.jnc.2018.02.009.

Nixon SW (1995) Coastal marine eutrophication: A definition, social causes, and future concerns. Ophelia 41(1):199–219. https://doi.org/10.1080/00785236.1995.10422044.

Nyshadham, C, Rupp M, Bekker B, Shapeev AV, Mueller T, Rosenbrock CW, Csányi G, Wingate DW, Hart GLW (2019) Machine-learned multi-system surrogate models for materials prediction. npj Computational Materials 1(5):51-56.

OECD Organisation for Economic Co- Operation and Development (1982): Eutrophication of Water. Monitoring, Assessment and Control. Environment Directorate, OECD, Paris. p 154.

Oh HM, Ahn CY, Lee JW, Chon TS, Choi KH, Park YS (2007) Community patterning and identification of predominant factors in algal bloom in Daechung Reservoir (Korea) using artificial neural networks. Ecol. Model. 203:109–118. https://doi.org/10.1 016/j.ecolmodel.2006.04.030.

Omer FD (2010) A hybrid neural network and ARIMA model for water quality time series prediction. Eng. Appl. Artif. Intell. 23:586–594.

Onderka M (2007) Correlations between several environmental factors affecting the bloom events of cyanobacteria in Liptovska Mara reservoir (Slovakia)-a simple regression model. Ecol. Model. 209:412–416. https://doi.org/10.1016/j.ecolmodel.2007.07.028.

Otto P, Day STA (2007) Biologist's Guide to Mathematical Modeling in Ecology and Evolution. Princeton University Press. Chapter Title: Mathematical Modeling in Biology. 1-14 pages. http://www.jstor.com/stable/j.ctvcm4hnd.4.

Ouchi A (1982) Forecast of a red tide occurrence in the northern part of Hiroshima Bay-II. Prediction of red tide occurrence by means of multiple linear regression model. Bull. Jpn. Soc. Sci. Fish. 48:1245–1250.

Oujidi B, Tahri M, Layachi M, Abid A, Bouchnan R, Selfati M, Bounakhla M, El Bouch M, Maanan M, Bazairi H, Snoussi M (2020) Effects of the watershed on the seasonal variation of the surface water quality of a post-restoration coastal wetland: The case of the Nador lagoon (Mediterranean sea, Morocco). Regional Studies in Marine Science 35:101-127. https://doi.org/10.1016/j.rsma.2020.101127.

Paerl HW, Huisman J (2008) Blooms like it hot. Science 320(5872):57-58. https://doi.org/10.1126/science.1155398.

Pai PF, Lin CS (2005) A hybrid ARIMA and support vector machines model in stock price forecasting. Omega 33:497–505. https://doi.org/10.1016/j.omega.2004.07.024.

Pajuelo JG, Lorenzo JM (1995) Analysis and forecasting of the demersal fishery of the Canary Islands using an ARIMA model. Sci Mar 59:155–164.

Palani S, Liong SY, Tkalich P (2008) An ANN application for water quality forecasting. Mar. Pollut. Bull. 56:1586–1597.

Park YS, Cereghino R, Compin A, Lek S (2003) Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. Ecol. Model. 160: 265–280.

Park Y, Cho KH, Park J, Cha SM, Kim JH (2015) Development of early-warning protocol for predicting chlorophyll-a concentration using machine learning models in freshwater

and estuarine reservoirs, Korea. Science Total Environment 502: 31-41. https://doi.org/10.1016/j.scitotenv.2014.09.005.

Parsons TR, Maita Y, Lalli CM (1984) A manual of chemical and biological methods for seawater analysis. Pergamon, Oxford sized algae and natural seston size fractions. Pergamon Press, Oxford Oxfordshire, New York.

Patterson S (2010). "Letting the Machines Decide". The Wall Street Journal.

Pearson MJ, Bateman IJ, Codd GA (2001) Measuring the recreational and amenity values affected by toxic cyanobacteria. A Contingent Valuation Study of Rutland Water, Leicestershire. Springer, Netherlands: pp.67–89.

Pelikan E, de Groot C, Wurtz D (1992) Power consumption in West-Bohemia: improved forecasts with decorrelating connectionist networks. Neural Network World 2:701–712.

Prista N, Diawara N, Costa MJ, Jones C (2011) Use of SARIMA models to assess data-poor fisheries: a case study with a sciaenid fishery of Portugal. Fish Bull 109(2):170–185.

Peters GH (1991) Agriculture and the macro-economy: presidential address. Journal of Agricultural Economics 42(3):231-249. https://doi.org/10.1111/j.1477-9552.1991.tb00354.x.

Pérez-Ruzafa A, Mompeán MC, Marcos C (2007) Hydrographic, geomorphologic and fish assemblage relationships in coastal lagoons. In: Viaroli P., Lasserre P., Campostrini P. (eds) Lagoons and Coastal Wetlands in the Global Change Context: Impacts and Management Issues. Developments in Hydrobiology, vol 192. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-6008-3_10.

Pérez-Ruzafa A, Campillo S, Fernández-Palacios JM, García-Lacunza A, García-Oliva M, Ibañez H, Navarro-Martínez PC, Pérez-Marcos M, Pérez-Ruzafa IM, Quispe-Becerra JI, Sala-Mirete A, Sánchez O and Marcos C (2019) Long-Term Dynamic in Nutrients, Chlorophyll a, and Water Quality Parameters in a Coastal Lagoon During a Process of Eutrophication for Decades, a Sudden Break and a Relatively Rapid Recovery. Front. Mar. Sci. 6:26. doi: 10.3389/fmars.2019.00026.

Phillips G, Pietiläinen O, Carvalho L et al (2008) Chlorophyll–nutrient relationships of different lake types using a large European dataset. Aquat Ecol 42: 213–226. https://doi.org/10.1007/s10452-008-9180-0.

Piante C, Ody D (2015) Blue Growth in the Mediterranean Sea: The Challenge of Good Environmental Status. Med Trends Project. WWF, France. p 89.

Pimienta J (1959) Le cycle pliocène-actuel dans les bassins paraliques de Tunis. Mémoires de la Société Géologique de France, Paris. p 176.

Plan Bleu (2016) Tourism and Sustainability in the Mediterranean: Key Facts and Trends. Plan Blue. Regional Activity Centre, Valbonne.

Prista N, Diawara N, Costa MJ, Jones C (2011) Use of SARIMA models to assess data-poor fisheries: a case study with a sciaenid fishery of Portugal. Fish Bull 109:170–185.

Rajaee T, Ebrahim H, Nourani F (2019) A review of the artificial intelligence methods in groundwater level modeling. Journal of Hydrology 572: 336-351. https://doi.org/10.1016/j.jhydrol.2018.12.037.

Raman RK, Mohanty SK, Bhatta KS, Karna SK, Sahoo AK, Mohanty BP, Das B.K (2018) Time series forecasting model for fisheries in Chilika lagoon (a Ramsar site, 1981), Odisha, India: a case study. Wetlands Ecol Manage 26: 677–687. https://doi.org/10.1007/s11273-018-9600-4.

Recknagel F, Ostrovsky I, Cao HQ, Zohary T, Zhang XQ (2013) Ecological relationships, thresholds and time-lags determining phytoplankton community dynamics of Lake Kinneret, Israel elucidated by evolutionary computation and wavelets. Ecol. Model. 255:70–86. https://doi.org/10.1016/j.ecolmodel.2013.02.006.

Recknagel F, Orr P, Cao H (2014) Inductive reasoning and forecasting of population dynamics of Cylindrospermopsis raciborskii in three sub-tropical reservoirs by evolutionary computation. Harmful Algae 31:26–34. https://doi.org/10.1016/j.hal.2013.09.004.

Rezgui A, Ben Maiz N, Moussa M (2008) Fonctionnement hydrodynamique et écologique du Lac Nord de Tunis par modélisation numérique. Revue des Sciences de l'Eau 21: 349-361. https://doi.org/10.7202/018781ar.

Robledano F, Esteve MA, Martínez-Fernández J, Farinós P (2011) Determinants of wintering waterbird changes in a Mediterranean coastal lagoon affected by eutrophication. Ecological Indicators 11:395-406. https://doi.org/10.1016/j.ecolind.2010.06.010.

Romilly P (2005) Time series modelling of global mean temperature for managerial decision-making. J Environ Manag 76:61–70.

Sahoo S, Jha MK (2013) Groundwater-level prediction using multiple linear regression and artificial neural network techniques: a comparative assessment. Hydrogeol. J. 21:1865–1887.

Sakka Hlaili A, Grami B, Niquil N, Gosselin M, Hammel D, Trousselier M, Hadj Mabrouk H (2008) The planktonic food web of the Bizerte lagoon (south-western Mediterranean) during summer: I. Spatial distribution under different anthropogenic pressures. Estuarine, Coastal and Shelf Science 78:71-77

Samarasinghe S (2007) Neural Networks for Applied Sciences and Engineering. Auerbach Publications, New York. p 570.

Satta CT, Pulina S, Rene A, Padedda BC, Caddeo T, Fois N, Luglie A (2020) Ecological, morphological and molecular characterization of Kryptoperidinium sp. (Dinophyceae) from two Mediterranean coastal shallow lagoons. Harmful Algae 97:101855. https://doi.org/10.1016/j.hal.2020.101855.

Schlink U, Doring S, Pelikan E et al (2003) A rigorous inter-comparison of ground-level ozone predictions. Atmos. Environ. 37:3237–3253.

Schneider W (1977) Bilan des substances nutritives du lac de Tunis 1976/1977. Rapp. Inst. Frésénius Gmbh (R.F.A), Minist.de l'Agriculture, Tunisie. p. 1-787.

Schramm W (1999) Factors influencing seaweed responses to eutrophication: some results from EU-project EUMAC. J. Appl. Phycol 11: 69–78.

Schwartz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464.

Sfriso A, Facca C, Ghetti PF (2003) Temporal and spatial changes of macroalgae and phytoplankton in a Mediterranean coastal area: the Venice lagoon as a case study. Marine Environmental Research 56, 617-636.

Sharda R, Patil RB (1990) Neural networks as forecasting experts: an empirical test, in: Proceedings of the International Joint Conference on Neural Networks, Washington, D.C., Vol. 2, pp. 491–494.

Shili A (1995) Contribution à l'étude de Ruppia dans le Lac Nord de Tunis. Mémoire de Diplôme des Etudes Approfondies en biologie marine et océanographie, Faculté des Sciences de Tunis, p.1- 128.

Shili A, Trabelsi EB, Ben Maiz N (2002) Seasonal dynamics of macroalgae in the South Lake of Tunis. Journal of Coastal Conservation 8: 127-134.

Shili A (2008) Les peuplements à Ruppia (Monocotyledone, Ruppiaceae) des milieux lagunaires de Tunisie. PhD thesis. Univ. Méd, p 305.

Singh N, Singh BK, Sinha N, Kumar B (2009) A response to comment on the paper "Growth and characterization of new nonlinear optical thiourea l-alanine acetate single crystal" Neelam Singh et al [J. Crystal Growth 310 (2008) 4487–4492]. Journal of Crystal Growth 311:1385–1386. https://doi.org/10.1016/j.jcrysgro.2008.12.045.

Song K, Park YS, Zheng F, Kang H (2013) The application of artificial neural network (ANN) model to the simulation of denitrification rates in mesocosm-scale wetlands. Ecol. Inform. 16: 10–16.

Souchu P, Bec B, Smith VH, Laugier T, Fiandrino A, Benau L, Orsoni V, Collos Y, Vaquer A (2010) Patterns in nutrient limitation and chlorophyll-a along an anthropogenic eutrophication gradient in French Mediterranean coastal lagoons. Canadian Journal of Fisheries and Aquatic Sciences 67:743-753. https://doi.org/10.1139/F10-018.

Srichandan S, Ji Yoon K, Kumar A, Mishra Deepak R, Bhadury P, Muduli Pradipta R, Pattnaik A, Rastogi G (2015) Interannual and cyclone-driven variability in phytoplankton communities of a tropical coastal lagoon. Marine Pollution Bulletin 101:39-52. https://doi.org/10.1016/j.marpolbul.2015.11.030.

Stergiou KI (1991) Describing and forecasting the sardine-anchovy complex in the eastern Mediterranean using vector autoregressions. Fish Res 11:127–141.

Stergiou KI, Chritou ED, Petrakis G (1997) Modelling and forecasting monthly fisheries catches: comparison of regression, univariate and multivariate time series methods. Fish Res 29:55–95.

Strickland JDH, Parsons TR (1972) A Practical Handbook of Seawater Analysis. Ottawa: Fisheries Research Board of Canada, Bulletin 167, 1968. 293 pp.

Strobl C, Boulesteix AL, Augustin T (2006) Unbiased split selection for classification trees based on the Gini index. Computational Statistics & Data Analysis 52:483-501. https://doi.org/10.1016/j.csda.2006.12.030.

Strobl C, Boulesteix AL, Kneib T, Augustin Zeileis TA (2008) Conditional variable importance for random forests. BMC Bioinformatics 9:307. https://doi.org/10.1186/1471-2105-9-307.

Su J, Wang X, Zhao S, Chen B, Li C, Yang Z (2015) A structurally simplified hybrid model of genetic algorithm and support vector machine for prediction of Chlorophyll a in reservoirs. *Water 7*(4): 1610-1627. https://doi.org/10.3390/w7041610.

Sujitha R, Ravindhranath K (2017) Extraction of Phosphate from Polluted Waters Using Calcium Alginate Beads Doped with Active Carbon Derived from A. aspera Plant as Adsorbent. Journal of Analytical Methods in Chemistry. Volume 2017, Article ID 3610878, 13 pages. https://doi.org/10.1155/2017/3610878.

Sun H, Koch M (2001) Case study: analysis and forecasting of salinity in Apalachicola Bay, Florida, using Box-Jenkins ARIMA models. J Hydraul Eng 127(9):718–727. https://doi.org/10.1061/(ASCE)0733-9429(2001)127:9(718).

Tang Z, Almeida C, Fishwick PA (1991) Time series forecasting using neural networks vs Box–Jenkins methodology, Simulation 57 : 303–310.

Tang Z, Fishwick PA (1993) Feedforward neural nets as models for time series forecasting, ORSA J. Comput. 5:374–385.

Thornber B, Mosedale A, Drikakis D, Youngs D, Williams RJR (2008) An improved reconstruction method for compressible flows with low Mach number features. Journal of Computational Physics 227:4873–4894. http://dx.doi.org/10.1016/ j.jcp.2008.0 1.036.

Tian W, Liao Z, Zhang J (2017) An optimization of artificial neural network model for predicting chlorophyll dynamics. Ecological Modelling 364: 42-52. http://dx.doi.org/10.1016/j.ecolmodel.2017.09.013.

Tiao GC (2001) Time Series: ARIMA Methods, Computer Science 363–407.

Tiyasha T, Tung.M T, Zaher Mundher Y (2020) A survey on river water quality modelling using artificial intelligence models: 2000–2020. Journal of Hydrology 585:124670. https://doi.org/10.1016/j.jhydrol.2020.124670.

Tizro AT, Ghashghaie M, Georgiou P, Voudouris K (2014) Time series analysis of water quality parameters. Journal of Applied Research in Water and Wastewater 1(1):40-50.

Trabelsi-Annabi N (2001) Contribution à l'étude de l'environnement du lac Nord de Tunis, répartition et structure des peuplements planctoniques. Mémoire de Diplôme des Etudes Approfondies en écologie générale, faculté des sciences de Sfax.162 p.

Trabelsi Bahri EL, Armi Z, Trabelsi Annabi N, Shili A, Ben Maiz N (2013) Water quality variables as indicators in the restoration impact assessment of the north lagoon of Tunis, South Mediterranean. Journal of Sea Research 79:12-19. http://dx.doi.org/10.1016/j.seares.2013.01.003.

Turki S, Balti N, Ben Salah C (2007) First detection of Kryptoperidinium foliaceum (Stein1883) in Tunisian waters. Harmful Algae News 35:9–10.

UNEP/MAP (2012) State of the Mediterranean marine and coastal environment. UNEP/ MAP – Barcelona Convention. Athens.

Van Berk H, Oostinga H (1992) North Lake of Tunis and its shores: restoration and development. Terra et Aqua 49:23–32.

Viaroli P, Christian RR (2003) Description of trophic status, hyperautotrophy and dystrophy of a coastal lagoon through a potential oxygen production and consumption index—TOSI: Trophic Oxygen Status Index. Ecological Indicators 3(4):237-250. DOI: 10.1016/j.ecolind.2003.11.001.

Viaroli P, Bartoli M, Azzoni R, Giordani G, Mucchino M, Naldi M, Nizzoli D, Taje L (2005) Nutrient and iron limitation to Ulva blooms in a eutrophic coastal lagoon (Sacca di Goro, Italy). Hydrobiologia 550:57–71. http://dx.doi.org/10.1007/s10750-005-4363-3.

Viaroli P, Bartoli M, Giordani G, Naldi M (2008) Community shifts, alternative stable states, biogeochemical controls and feedbacks in eutrophic coastal lagoons : a brief overview. Aquat. Conserv. Mar. Freshw. Ecosyst. 18: 105–117. https://doi.org/ 10.1002/aqc.

Villanoy CL, Azanza RV, Altemerano A, Casil AL (2005) Attempts to model the bloom of Pyrodinium, a tropical toxic dinoflagellate. Harmful Algae 5:156–183. https://doi.org/10.1016/j.hal.2005.07.001.

Vogiatzakis IN, Mannion AM, Griffiths GH (2006) Mediterranean ecosystems: problems and tools for conservation. Progress in Physical Geography 30(2):175–200. https://doi.org/10.1191/0309133306pp472ra.

Wallace J, Champagne P, Hall G (2016) Time series relationships between chlorophyll-a, dissolved oxygen, and pH in three facultative wastewater stabilization ponds. Environ. Sci.: Water Res. Technol. 2:1032-1040. https://doi.org/10.1039/C6EW00202A.

Wang Z, Huang K, Zhou PJ, Guo HC (2010) A hybrid neural network model for cyanobacteria bloom in Dianchi Lake. Procedia Environ. Sci. 2:67–75.

Water Framework Directive, 2000. Directive 2000/60/EC (WFD). OJL 327. pp. 1–73

Watzin MC, Miller EB, Shambaugh AD, Kreider MA (2006) Application of the WHO alert level framework to cyanobacterial monitoring of Lake Champlain, Vermont. Environmental Toxicology 21:278-288. https://doi.org/ 10.1002/tox.20181.

Wei WWS (1990) Time series analysis: univariate and multivariate methods. 2th ed. Addison-Wesley Publishing Company, New York. 634 pp.

Wei B, Sugiura N, Maekawa T (2001) Use of artificial neural network in the prediction of algal blooms. Water Res. 35:2022–2028.

Wei WWS (2006) Time Series Analysis: Univariate and Multivariate Methods, 2nd Edn. London: Pearson. p 614.

Were K, Bui DT, Dick ØB, Singh BR (2015) A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. Ecological Indicators 52: 394-403. https://doi.org/10.1016/j.ecolind.2014.12.028.

Wheeler SM, Morrissey LA, Levine SN, Livingston GP, Vincent WF (2012) Mapping cyanobacterial blooms in Lake Champlain's Missisquoi Bay using Quick Bird and MERIS satellite data. Journal of Great Lakes Research 1: 68-75. https://doi.org/10.1016/j.jglr.2011.06.009.

World Health Organization & International Programme on Chemical Safety. (1996). Guidelines for drinking-water quality. Vol. 2, Health criteria and other supporting information, 2nd ed. World Health Organization. p 973. https://apps.who.int/iris/handle/10665/38551.

Wilhelm SW, Farnsley SE, LeCleir GR (2011) The relationships between nutrients, cyanobacterial toxins and the microbial community in Taihu (Lake Tai), China. Harmful Algae 10:207–215. https://doi.org/10.1016/j.hal.2010.10.001.

Williams JB, Poff NL (2006) Informatics software for the ecologist's toolbox: A basic example. Ecological Informatics 1: 325-329. https://doi.org/10.1016/j.ecoinf.2006.03.007.

Wu CL, Chau KW (2011) Rainfall-runoff modeling using artificial neural network coupled with singular spectrum analysis. Journal of Hydrology 399: 394-409. https://doi.org/10.1016/j.jhydrol.2011.01.017.

Xu G, Schwarz P, Yang H (2019) Determining China's CO2 emissions peak with a dynamic nonlinear artificial neural network approach and scenario analysis. Energy Policy 128: 752–762. https://doi.org/10.1016/j.enpol.2019.01.058.

Yi HS, Park S, An KG, Kwak KC (2018) Algal Bloom Prediction Using Extreme Learning Machine Models at Artificial Weirs in the Nakdong River, Korea. International Journal of Environmental Research and Public Health 15: 2078. https://doi.org/10.3390 /ijerph15102078.

Yu L, Liang S, Chen R, Lei KK (2019) Predicting monthly biofuel production using a hybrid ensemble forecasting methodology. International Journal of Forecasting https://doi.org/10.1016/j.ijforecast.2019.08.014.

Zaldívar JM, Cardoso AC, Viaroli P, de Wit R, Ibanez C, Reizopoulou S, Razinkovas A, Basset A, Holmer M, Murray N (2008a) Eutrophication in transitional waters: an overview. Transiti. Waters Monogr. 1:1–78. https://doi.org/10. 1285/i18252273v2n1p1.

Zaldívar JM, Strozzi F, Dueri S, Marinov D, Zbilut JP (2008b) Characterization of regime shifts in environmental time series with recurrence quantification analysis. Ecol. Model. 210:58–70. https://doi.org/10.1016/j.ecolmodel.2007.07.012.

Zang C, Huang S, Wu M et al (2010) Comparison of Relationships Between pH, Dissolved Oxygen and Chlorophyll a for Aquaculture and Non-aquaculture Waters. Water Air Soil Pollut 219:157–174. https://doi.org/10.1007/s11270-010-0695-3.

Zaouali J (1974) Etude de l'écologie du lac Nord de Tunis et de la mer de Bou Ghrara. Leurs peuplements macrologiques. Thèse 3éme cycle. Univ. CAEN, 345 p.

Zaouali J (1977) Le lac de Tunis : facteurs climatiques, physico-chimiques et crises dystrophie. Bull. Off. Nat. Pêche. Tunisie, 1 (1) : 37-49.

Zaouali J (1983) Historique de l'évolution sédimentologique de la lagune de Tunis. Service Géolo. De Tunisie, 47: 61-66.

Zaouali J, Baetten S (1984) Etude historique et évaluation actuelle de l'impact de l'eutrophisation sur l'écosystème de la lagune de Tunis. VIIèmes journées d'études sur les pollutions, lucernes, CIESM. (1984) : 671-678.

Zaouali J, Baetten S (1985) Etude historique et évaluation actuelle de l'impact de l'eutrophisation sur l'écosystème de la lagune de Tunis. In: Rapp. P.V. Réun. Commiss. Intern. Explor. Sci. Médit (29):671–678.

Zhang G, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: The state of art. International Journal of Forecasting 14 (1): 35-62.

Zhang GP, Patuwo EB, Hu MY (2001) A simulation study of arti/cial neural networks for nonlinear time-series forecasting, Comput. Oper. Res. 28: 381–396.

Zhang GP (2003) Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50:159-175.

Zhang GP, Qi M (2005) Neural network forecasting for seasonal and trend time series. Eur J Oper Res 160(2):501-514.

Zhang H, Hu W, Gu K, Li Q, Zheng D, Zhai S (2013) An improved ecological model and software for short-term algal bloom forecasting. Environ. Model. Softw. 48:152–162. https://doi.org/10.1016/j.envsoft.2013.07.001.

Zhang WZ, Wang H, Chai F, Qiu G (2016) Physical drivers of chlorophyll variability in the open South China Sea. Journal of Geophysical Research Oceans 121:7123-7140. https://doi.org/10.1002/2016JC011983.

**Electronic references**

[1] https://fr.mathworks.com/help/matlab/index.html

[2] https://netflixtechblog.com/netflix-recommendations-beyond-the-5-stars-part-1-55838468f429

[3] https://techcrunch.com/2012/01/10/doctors-or-algorithms/

[4] https://neptune.ai/blog/the-ultimate-guide-to-evaluation-and-selection-of-models-in-machine-learning

[5] https://techvidvan.com/tutorials/artificial-neural-network/

[6] https://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

[7] https://machinelearningmastery.com/time-series-data-Stationary-python/

[8] https://data.library.virginia.edu/understanding-q-q-plots/

**Résumé du projet de thèse en Français**

À l'échelle mondiale, les zones côtières sont gravement touchées par les activités anthropiques, ce qui fait de l'eutrophisation des écosystèmes côtiers un problème mondial, en particulier dans les lagunes (Nixon, 1995 ; Cloern, 2001).

L'augmentation des apports en nutriments, renforcée par l'urbanisation, l'agriculture ou l'industrie, conduit à une réponse complexe directe et/ou indirecte des écosystèmes naturels (Schramm, 1999 ; Viaroli *et al*., 2008), notamment des crises anoxiques, des proliférations d'algues toxiques, voire une perte d'espèces, et plus largement, la détérioration des fonctions des écosystèmes et des services qu'ils fournissent (Cloern, 2001 ; Zaldivar *et al*., 2008 a, b).
Les lagunes côtières méditerranéennes ont été exposées à l'eutrophisation anthropique depuis des décennies et font partie des systèmes les plus vulnérables à de telles pressions (Viaroli *et al*., 2005 ; Zaldívar *et al*., 2008 ; Souchu *et al*., 2010). Elles sont influencées par des côtes densément peuplées, en particulier pendant l'été, lorsque la Méditerranée devient la principale destination du tourisme mondial (Vogiatzakis *et al*., 2006). En plus de cette pression urbaine, l'export limité des eaux de ces écosystèmes vers la mer les rendent particulièrement vulnérables à l'eutrophisation (de Jonge and Elliott, 2001 ; Newton *et al*., 2014). Plusieurs études ont été menées dans les lagunes côtières méditerranéennes pour évaluer le niveau d'eutrophisation. García-Ayllón (2017), a déclaré que la lagune de Mar Menor, située à l'est de la région de Murcie en Espagne, a subi un processus d'anthropisation intense au cours des cinq dernières décennies. L'un des principaux indicateurs de cela était la croissance exponentielle de la population d'une nouvelle espèce de méduse atteignant plus de 100 millions d'individus chaque été (Robledano *et al*., 2011). La lagune de Thau est un autre cas particulièrement intéressant pour l'étude de l'eutrophisation des lagunes côtières méditerranéennes. C'est une lagune située sur la côte méditerranéenne française qui est célèbre pour ses activités traditionnelles de conchyliculture. Cette lagune a été soumise à une eutrophisation conduisant à des événements anoxiques majeurs associés à des mortalités massives de stocks de mollusques et crustacés (Derolez *et al*., 2020). On peut aussi citer le complexe lagunaire Palavasien, qui est un rassemblement de huit lagunes le long de la côte méditerranéenne française, qui a subi une eutrophisation intensive pendant quatre décennies principalement liée au sur-enrichissement en éléments nutritifs des rejets continus des eaux usées (Leruste *et al*., 2016). La lagune de Ghar El Melh fournit un bon exemple pour l'étude du processus d'eutrophisation dans les lagunes côtières méditerranéennes. Selon Shili *et al*. (2002), la lagune a connu plusieurs crises dystrophiques au cours de la période allant de 1994

jusqu'à 1996. En plus, Turki *et al*. (2007) ont signalé la prolifération d'espèces d'algues nuisibles dans la lagune, telles que *Kryptoperidinium foliaceum*, *Prorocentrum micans* et *Anabaena sp*.

En vue de tout ça, la Lagune Nord de Tunis, une lagune sud méditerranéenne située au nord de la Tunisie (Fig. 1), fournit un bon exemple pour le diagnostic et l'étude de l'eutrophisation dans les écosystèmes côtiers méditerranéens. En effet, la Lagune Nord de Tunis est l'une des lagunes les plus importantes de Tunisie, qui a connu un état écologique critique essentiellement dû au développement urbain (Harbridge *et al*., 1976; Zaouali et Baetten, 1984; Zaouali et Baetten 1985). En effet, les changements radicaux et très apparents affectant la biodiversité ont révélé la gravité du problème écologique de ce milieu lagunaire poussant ainsi l'homme à intervenir afin d'améliorer une telle situation critique. Plusieurs études sur le milieu ont mené à la mise en place d'un projet d'assainissement qui a pu être concrétisé sous la direction de la Société Al Buhaira de développement et d'investissement en 1985.

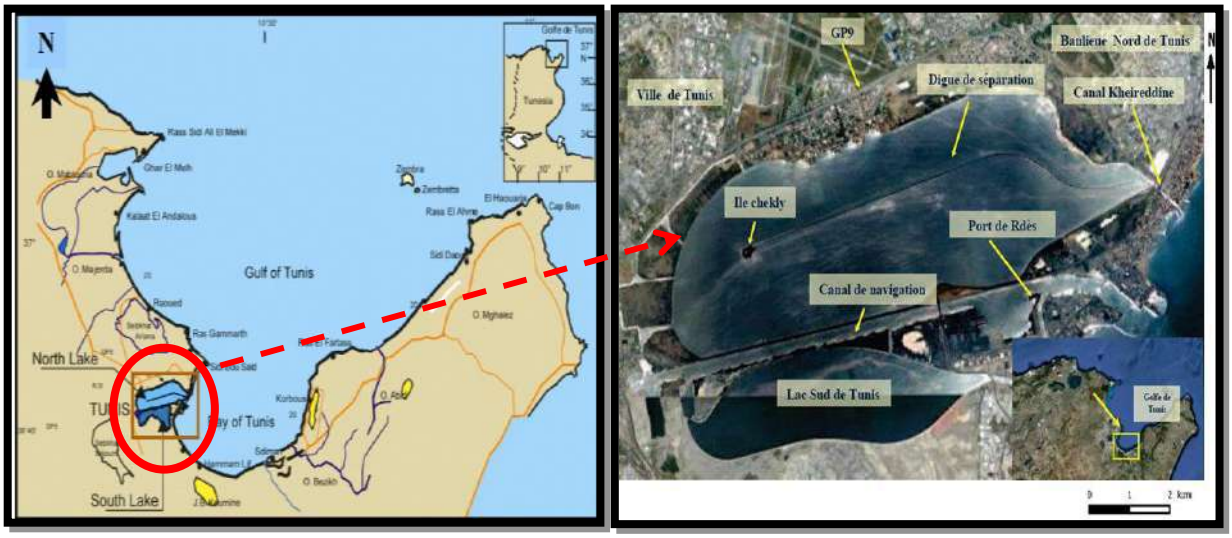



**Fig 1. Localisation de la zone d'étude**

Les changements et les modifications observés à la suite de ce projet de restauration sont marqués par une nette amélioration de la qualité des eaux du Lac et de la biodiversité (Ben Maiz, 1997 ; Shili, 2008). La Lagune Nord de Tunis est un milieu très productif en Tunisie. C'est une zone importante notamment pour les oiseaux migrateurs. Cette lagune a été nommée zone humide d'importance internationale (site Ramsar) en 2013.

Dans ce meme contexte, la mesure de la chlorophylle-*a* (Chl-*a*) est utilisée comme indicateur de la biomasse phytoplanctonique présente dans l'eau et donc du degré d'eutrophisation du milieu. C'est le pigment le plus important chez les organismes photosynthétiques aérobies du

règne végétal (Tian *et al*., 2017). Des niveaux élevés de Chl-*a* pourraient être interprétés comme indicateur de proliférations d'algues, ce qui a des effets importants sur les processus physiques, chimiques et biologiques d'une lagune. Les Cyanotoxines générées par les cyanobactéries dans l'eau des lagunes pourraient présenter un risque pour la santé humaine (Watzin *et al*., 2006 ; Mc Quaidet *et al*., 2011 ; Kalaji *et al*., 2016). Dans le cas où la concentration en cyanotoxines dans un écosystème donné n'est pas disponible, la Chl-*a* est également largement acceptée comme mesure indirecte de la densité des cyanobactéries (Wheeler *et al*., 2012). Par conséquent, il est essentiel de contrôler les concentrations en Chl-*a* pour fournir des informations concernant le degré d'eutrophisation d'un écosystème naturel donné.

La possibilité de surveiller d'une manière automatique la qualité de l'eau, avec les modèles prédictifs, est très utile, en particulier dans les zones sensibles où (1) la menace d'épisodes potentiels de pollution est élevée et (2) des activités socio-économiques pertinentes, qui nécessitent des actions préventives, sont réalisées. Cependant, pour autant que l'on sache, il n'y a pas de dispositif automatique qui mesure avec précision la Chl-*a* en temps réel. Les mesures de Chl-*a* doivent être effectuées en laboratoire, ce qui signifie des taches longues à réaliser avec des coûts élevés (Jimeno-Sáez *et al*., 2020). Pour éviter de tels désagréments, la plupart des écologistes se sont récemment mis à diverse techniques de modélisation.

Le monde change à un rythme sans précédent, ce qui rend la recherche sur les écosystèmes de plus en plus complexe et de nombreuses nouvelles questions émergent chaque jour (Griitzner, 1996). Certaines approches standard telles que le travail de terrain ou les expérimentations en laboratoires seules, ne sont plus suffisantes pour la description précise et complète des écosystèmes complexes (Griitzner, 1996). Depuis longtemps, l'écologie a reconnu la nécessité de travailler avec les mathématiques, les statistiques et les disciplines d'informatiques (Otto and Day, 2007). Cette approche interdisciplinaire est élégamment soulignée dans certains des articles les plus importants et les plus cités, écrits par les pionniers de l'écologie, tels que Fisher, Preston et MacArthur qui ont utilisé des approches mathématiques pour expliquer et analyser les observations écologiques (Fisher *et al*., 1943 ; Mac Arthur, 1967). Les disciplines de l'écologie mathématique et théorique sont devenues un outil à utiliser pour les écologistes. Selon Codling and Dumbrell (2012), sans ces nouvelles disciplines de l'écologie, la collecte de données est une tâche futile et dénuée de sens.

Un modèle efficace relie les données à des questions écologiques, et fournit une part suffisante de compréhension ou de prévisions, là où la perception d'un écologiste est à elle seule insuffisante (Kompare *et al*., 1994). En termes de développement de modèles

écologiques, il existe généralement deux approches : (1) Physique (ou conceptuelle) et (2) basée sur les données (Babovic, 2005 ; He *et al*., 2014 ; Zhang *et al*., 2016). Les approches de modélisation de la qualité de l'eau fondées sur la physique sont capables de simuler les processus physiques internes du système aquatique, mais nécessitent des informations détaillées qui ne sont pas facilement accessibles (Dogan *et al*., 2009). De plus, de nombreux modèles physiques portant sur la qualité de l'eau prennent du temps à être compiler (Singh *et al*., 2009). Alors que les modèles basés sur les données sont plus simples à mettre en œuvre, pas si complexes et ils évitent le besoin de connaissances spécialisées des processus physiques (Bowden *et al*., 2006). Ces modèles sont populaires et couramment utilisés pour modéliser des processus naturels complexes, principalement dans la modélisation prédictive, car la prédiction du futur est considérée par beaucoup comme l'objectif ultime en écologie (Peters 1991). Malgré le nombre d'études qui se sont concentrées sur ce processus d'eutrophisation dans les écosystèmes aquatiques naturels (ex. Delbaere and Nieto-Serradilla, 2004 ; García Pintado *et al*., 2007), il y a encore relativement peu de travaux réalisés pour mettre en place une approche d'alerte proactive, pour identifier et prévenir les problèmes potentiels en avance, notamment à l'échelle méditerranéenne.

Dans la modélisation prédictive, différentes approches statistiques ont été utilisées basée sur les analyses de régression. Cependant, ces méthodes de traitement de données appliquent généralement une relation linéaire pour simplifier des problèmes complexes, ce qui conduit à des résultats insatisfaisants car ils ne sont pas suffisamment efficaces pour traiter des relations non linéaires compliquées entre les variables impliquées (Su *et al*., 2015). Une méthodologie productive et rentable sur l'estimation de l'eutrophisation avec des approches avancées, efficaces et robustes est nécessaire pour atteindre l'objectif d'une prédiction proche de la réalité (Tiyasha *et al*., 2020), et surmonter les limitations évoquées précédemment.

Les algorithmes d'apprentissage automatique (Machine Learning ou ML) se sont révélés plus efficaces, que les approches traditionnelles de traitement des données pour déterminer la qualité de l'eau (Abba *et al*., 2017), car ils sont très bien adaptés pour traiter des fonctions non linéaires et complexes. Des études antérieures ont confirmé la supériorité de ML sur les approches traditionnelles dans la modélisation des paramètres de la qualité de l'eau (Juntunen *et al*., 2012 ; Charulatha *et al*., 2017).

Les principales techniques de modélisation prédictives utilisées dans cette étude, pour prédire les concentrations de la Chl-*a*, en tant qu'indicateur de l'eutrophisation, dans la Lagune Nord de Tunis, sont brièvement décrites, ainsi que les résultats obtenus.

Dans ce projet de thèse, le premier chapitre a été consacré à la présentation du milieu d'étude et à la caractérisation spatio-temporelle et saisonnière des paramètres physico-chimiques dans la Lagune Nord de Tunis. En effet, l'eutrophisation des lagunes côtières est un phénomène complexe en raison de la grande variabilité des conditions spatiales et temporelles (Coelho *et al*., 2015).

Pour se faire, nous avons utilisé une série chronologique mensuelle de janvier 1989 jusqu'à avril 2018 (environ trois décennies) pour chaque variable (Chlorophylle-*a*, profondeur, Azote total, Phosphore total, pH, salinité, oxygène dissous et température).

Les travaux de terrain, les analyses en laboratoire, l'étude bibliographique ainsi que la contribution de la Société Al Buhaira de développement et d'investissement, nous ont permis de reconstituer ces séries chronologiques. Ces paramètres ont été échantillonnés à cinq Stations représentatives de la lagune (Fig. 2), mises en place selon le modèle hydrodynamique réalisé pour le projet de restauration évoqué un peu plus haut.

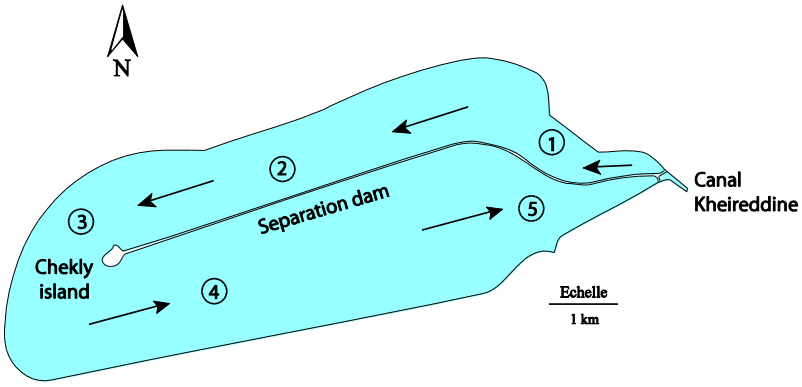



**Fig 2. Les Stations d'échantillonnage (1-5). Les flèches à l'intérieur de la lagune représentent le système de circulation unidirectionnelle de l'eau à l'entrée et à la sortie du milieu.**

La représentation spatio-temporelle et saisonnière de ces variables nous ont permis de conclure que :

La température de l'eau augmente progressivement à mesure que la température de l'air ambiant augmente, mais nous n'avons pas signalé d'augmentation brutale, ce qui confirme l'absence de pollution thermique dans la zone d'étude. En effet, la lagune étudiée a fait l'objet d'une pollution thermique causée par les rejets de la Société Tunisienne de l'Electricité et du Gaz (STEG) avant le projet d'assainissement (Ben Maiz, 1992).

La salinité dépend de la position du lieu d'échantillonnage, la partie nord de la lagune est caractérisée par une concentration de salinité plus faible que la partie sud. En effet, les eaux marines du golfe de Tunis entrent dans la lagune par la partie nord. La stratégie de circulation de l'eau permet son chargement en nutriments et en matières en suspension pendant son temps de séjour dans la lagune, ce qui contribue à l'augmentation de la salinité dans la partie sud. De plus, dans la partie sud, l'évaporation est plus intense (faible profondeur), car, lors de la réalisation du projet de restauration, les travaux de dragage ont principalement concerné la partie nord de la lagune.

La teneur en salinité dépend également de la saison, en raison de l'influence des précipitations notamment aux Stations proches des zones où l'eau de pluie est évacuée (Stations 1,2 and 3).

Les valeurs de pH enregistrées ne dépassent pas les normes selon APHA (1999) et qui sont de nature basique (pH>7).

L'oxygène dissous est un paramètre essentiel dans le maintien de la vie aquatique, tels que les phénomènes de dégradation de la matière organique et la photosynthèse.

Dans la Lagune Nord de Tunis, les teneurs en oxygène dissous sont plus élevées du côté nord, où les eaux marines du golfe de Tunis entrent, plutôt que du côté sud, où les eaux relativement peu profondes sont chargée en nutriments ce qui réduit la teneur en oxygène dissous.

Concernant la composante biologique et les nutriments, certaines valeurs relativement élevées de la concentration en Chl-*a*, ainsi qu'en azote total et phosphore total, peuvent indiquer un état d'eutrophisation dans certaines zones de la lagune, en particulier dans la partie sud.

Parmi les données environnementales et biologiques, l'oxygène dissous, les concentrations de phosphore total, les concentrations de salinité et les valeurs de la Chl-*a* ont été les plus affectées par les variations saisonnières. Les valeurs les plus élevées ont été enregistrées pour la salinité, le phosphore total et l'azote total pendant la saison estivale (ou sèche). Au printemps, la Chl-*a*, suivi de l'oxygène dissous ont montré des valeurs élevées.

En général, la plupart des paramètres présentent une variation spatiale et saisonnière significative. Aussi, ils ont montré que les eaux dans la Lagune Nord de Tunis étaient fortement influencées par l'apport d'eau marine du golfe de Tunis.

D'une manière générale, la qualité physico-chimique de l'eau concorde assez bien avec l'amélioration marquée de l'état d'eutrophisation de la Lagune Nord de Tunis. Cependant,

compte tenu de la fragilité de cet écosystème, il reste très influencé notamment par les conditions climatiques qui ont un effet direct sur la température, la Chl-*a* et la salinité.

Cet environnement doit rester sous surveillance et les composantes biotiques et abiotiques doivent être contrôlées régulièrement dans les années à venir.

Sachant cela, la modélisation prédictive semble être une technique appropriée pour garder la Lagune Nord de Tunis sous surveillance.

Afin de déterminer l'algorithme le plus approprié, le prétraitement des données comme cela a été fait dans le premier chapitre, est une étape essentielle pour capturer toutes spécificités cachées comme spatiales, temporelles, saisonnières, linéaires ou non linéaires. En effet, avant de se lancer dans la modélisation un peu complexe, il faut identifier et prendre en compte ces spécificités. Zhang et Qi (2005), ont créé un modèle prédictif qui tient compte de la saisonnalité, en soulignant la nécessité d'un prétraitement approprié des données pour tenir compte des fluctuations saisonnières ou des tendances de la série temporelle.

L'approche proposée dans le second chapitre, tiens compte de ces spécificités précédemment évoquées, et repose sur une combinaison de méthodes ML, utilisant en premier lieu la technique des forêts aléatoires (Random Forest ou RF) et en second lieu l'algorithme relatif aux réseaux artificiels de neurones appelé réseau de neurones autorégressif non linéaire avec des entrées externes (Nonlinear autoregressive with external inputs neural network ou NARX) pour prédire et prévoir les concentrations de la Chl-*a* dans la Lagune Nord de Tunis, en tant qu'indicateur d'eutrophisation.

Le NARX fait partie des algorithmes des réseaux artificiels de neurones (Artificial Neural Networks ou ANNs) qui sont très utilisés dans la modélisation prédictive.

Les ANNs imitent les processus d'apprentissage de l'être humain (le fonctionnement du cerveau), à travers la formation et le calibrage du réseau. Cette capacité fait des ANNs des outils précieux pour étudier des scénarios complexes non-linéaires, difficiles à décrire avec les méthodes conventionnelles (Daliakopoulos *et al*., 2005 ; Samarasinghe, 2006).

La capacité à saisir la dynamique du système et les non-linéarités rendent les ANNs exceptionnellement adéquats pour l'étude des systèmes naturels qui présentent généralement des variabilités spatio-temporelles distinctes (ASCE, 2000). Les algorithmes des ANNs ont également été appliquées à la dynamique de la Chl-*a*, car sa concentration s'agit de l'un des facteurs de représentation de la biomasse algale, et ont été considérées comme l'une des approches proactives d'alerte précoce pour prévenir l'apparition de certains épisodes de

proliférations d'algues (eutrophisation). Li *et al*. (2017) et Yi *et al*. (2018) ont appliqué différents types d'ANNs pour estimer la concentration de la Chl-*a* dans 27 lacs en Chine et dans une rivière en Corée, respectivement. Tian *et al*. (2017) ont utilisé un ANN pour prédire les concentrations de la Chl-*a* dans un réservoir d'estuaire en Chine orientale.

Il existe également de nombreuses classifications dans les ANNs, tels que les réseaux de propagation arrière (back-propagation networks), les réseaux de fonction de base radiale (Radial Basis Function networks) et les réseaux de Hopfield (Hopfield networks).

La propagation arrière est un algorithme d'apprentissage couramment utilisé dans les applications des ANNs. Le NARX est un réseau neuronal artificiel dynamique, formant un groupe important de réseaux neuronaux non linéaires à propagation arrière. Le NARX nous permet de prévoir le contenu de la Chl-*a* dans l'écosystème à court terme en se basant sur les données antérieures de la Chl-*a*, ainsi des entrées externes (autres variables en relation avec la Chl-*a*).

Pour pouvoir optimiser le temps de travail et réduire le cout, il est très important de limiter le nombre de paramètres à mesurer. C'est pourquoi, dans cette étude, il est très important de sélectionner les variables spécifiques qui sont les plus liées à la concentration de la Chl-*a*.

Pour se faire, la technique de ML appelée RF est une approche efficace. RF est une technique qui présente plusieurs avantages (Beriman, 2001). RF peut être formée sur de petits ensembles de données et également sur un grand nombre de variables prédictives, en choisissant parmi eux les plus utiles dans le cadre du champ d'application donnée (Park *et al*., 2015). Les prévisions de RF sont également très fiables car elles proviennent d'une moyenne d'ensemble de plusieurs forets simples, ce qui permet d'éviter le problème de sur-ajustement typique de nombreuses techniques de régression non linéaire (Huang *et al*., 2015).

Le modèle RF a été comparé à une régression linéaire multivariée (Multivariate Linear Regression ou MVLR) et a montré plus d'efficacité.

Le RF a été appliquée dans de nombreuses études. Béjaoui *et al*. (2016) ont étudié avec le modèle RF les variables prédictives les plus importantes pour la variation de la Chl-*a* dans la lagune de Bizerte située au nord de la Tunisie. Dans une autre étude plus récente, Béjaoui *et al*. (2018), ont utilisé le modèle RF pour étudier la dynamique du plancton dans la lagune de Ghar El Melh, située au nord de la côte méditerranéenne tunisienne.

Les objectifs de ce second chapitre sont les suivants: (1) sélectionner les variables spécifiques les plus liées à la production de la Chl-*a* en utilisant le modèle RF dans la Lagune Nord de

Tunis, pour se faire, plusieurs scénarios ont été testés (2) de développer un modèle prédictif pour estimer et prévoir avec une longueur d'avance la concentration de la Chl-*a* en utilisant le NARX et (3) pour valider la performance du modèle prédictif dans l'estimation de la concentration de la Chl-*a* pour la Lagune Nord de Tunis.

Pour s'assurer de la bonne sélection des variables les plus importantes à la variation de la Chl-*a*, plusieurs scénarios de NARX avec différentes combinaisons de variables (inputs) ont été examinés : sans sélection, sélection des trois variables les plus importantes selon le modèle RF, sélection des deux variables les plus importantes selon le modèle RF, et sélection des trois variables significativement corrélées à la Chl-*a* selon MVLR.

Le scénario avec les deux variables les plus importantes selon le modèle RF a été le plus performant.

RF a montré que la profondeur de Secchi ainsi que l'oxygène dissous sont essentiellement les principaux contributeurs à la variation de la Chl-*a*. Ces résultats concordent bien avec ceux issues d'autres études menées sur les lagunes côtières méditerranéennes.

Afin de réduire le coût de la surveillance *in situ* de l'environnement aquatique, outre le gain de temps, il est très important de réduire le nombre de variables lors de la modélisation prédictive. La profondeur et l'oxygène dissous sont des variables appropriées très faciles à mesurer, sans nécessiter d'analyses supplémentaires en laboratoire.

Le NARX développé a été capable de bien prédire la dynamique de la concentration de la Chl-*a* en utilisant un minimum de variables. Les entrées externes du NARX sont les variables qui contribuent le plus à la variation de la Chl-*a* selon le modèle RF (la profondeur et l'oxygène dissous).

On peut utiliser ce modèle, surtout lorsque les données sur la Chl-*a* ne sont pas disponibles. Nos résultats montrent que le phénomène de l'eutrophisation a pu être modélisé en utilisant la technique du NARX et, certaines valeurs extrêmes ont été estimées avec succès.

La performance du NARX a été testée par des indices mathématiques. L'erreur quadratique moyenne (Mean Squared Error ou MSE) est une mesure caractérisant la « précision » d'un modèle donné. Elle peut être sensible aux valeurs aberrantes et elle est mieux utilisée en conjonction avec d'autres métriques pour évaluer un modèle donné lorsque des valeurs aberrantes sont présentes. Si le MSE est proche de 0, cela indique une approximation très proche des valeurs réelles.

La précision de la prédiction R représente le degré de corrélation entre les valeurs de prédiction et les valeurs observées, et une valeur R élevée proche de 1 signifie que la prédiction est proche de la valeur observée.

En outre, la fonction d'auto-corrélation des erreurs (the error autocorrelation function) et la fonction d'inter-corrélation des erreurs et des entrées (the input error cross correlation function) ont également été vérifiées pour évaluer la performance du NARX. La fonction d'auto-corrélation des erreurs décrit comment les erreurs de prédiction sont liées dans le temps. Pour un modèle de prédiction parfait, la différence entre les deux erreurs doit être suffisamment faible pour être statistiquement non significative. Cela signifie que les erreurs de prévision ne sont pas du tout corrélées. Cela indique que les valeurs de l'auto-corrélation des erreurs devraient se situer pour la plupart dans un certain intervalle de confiance, par exemple 95% (Xu *et al*., 2019). La fonction d'inter-corrélation des erreurs et des entrées indique comment les erreurs sont corrélées avec les séries temporelles d'entrée. Pour un modèle de prédiction parfait, toutes les corrélations doivent être nulles (dans un certain intervalle de confiance, par exemple 95%), sauf celle située au zéro décalage ou « zero lag » (Markova, 2019).

Le NARX réalisé a un R= 0.79 et un MSE= 0.31. Ce qui prouve que les prédictions de notre approche sont suffisamment robustes.

Il est important de mentionner que malgré la quantité importante de données observées (environ trois décennies) utilisées pour le développement du NARX, celui-ci a un temps de compilation très court.

Dans un contexte plus large sur les études qui portent sur les lagunes côtières et autres écosystèmes côtiers, notre approche pourrait être utilisée pour évaluer et prévoir le processus d'eutrophisation de ces milieux naturels et aider à la prise de décision des autorités civiles, ainsi que des ingénieurs, économistes, investisseurs et autres parties intéressées.

Le troisième chapitre porte sur la prédiction de la Chl-*a* en tant qu'indicateur de l'eutrophisation sur le long terme dans la Lagune Nord de Tunis, en utilisant la technique de Box et Jenkins (1967). Box et Jenkins, est une procédure itérative complexe, qui produit un modèle de moyenne mobile intégrée autorégressive, qui s'ajuste aux facteurs saisonniers et tendanciels (Gaynor and Kirkpatrick, 1994). La méthode est appropriée pour les séries chronologiques de moyenne à longue durée, au moins 50 observations sont nécessaires (Wei, 1990).

La méthode de Box et Jenkins a été sélectionné basée sur : 1) sa capacité à traiter des situations complexes ; 2) son adaptabilité dans le traitement de données de séries chronologiques dépendantes ; 3) ses processus mathématiques et statistiques avancés ; 4) sa fonctionnalité en matière d'analyse d'incertitude ; et 5) la simplicité de sa mise en œuvre.

Les méthodologies de Box et Jenkins produisent souvent les modèles de prévision les plus précis pour tout ensemble de données (Gaynor and Kirkpatrick, 1994). Ces méthodes permettent également une approche plus systématique de la construction, de l'analyse et de la prévision des modèles de séries chronologiques. Le test comparatif d'Armstrong (Armstrong, 1985) sur le classement des méthodes d'extrapolation (du rang le plus élevé "1" au rang le plus bas "5") en termes de coût, de compréhensibilité et de précision des prévisions, tant à court terme qu'à long terme a classé la méthode de Box et Jenkins comme 1,5 pour la précision des prévisions à court terme et 2 pour la précision des prévisions à long terme. En général, les méthodologies Box et Jenkins utilisent les observations les plus récentes comme valeurs de départ, puis analysent les erreurs de prévisions récentes afin de déterminer les ajustements appropriés pour les périodes futures. En faisant cela, elles permettent des ajustements efficaces des niveaux d'erreur et fournissent une imitation plus souple d'une tendance complexe particulière ou de la saisonnalité. Les modèles de Box et Jenkins sont capables de traiter les données de séries chronologiques dépendantes qui ne sont pas considérées comme adaptées à d'autres méthodes. Par exemple, un modèle de régression repose sur l'hypothèse standard que le terme d'erreur doit être statistiquement indépendant. En réalité, de nombreuses données liées au temps sont dépendantes ou corrélées entre elles (Lu *et al*., 2008), en d'autres mots, non Stationnaires. L'approche de Box et Jenkins, part de l'hypothèse que le processus qui a généré la série chronologique peut être approximé en utilisant un modèle ARMA s'il est Stationnaire ou un modèle ARIMA s'il est non Stationnaire (Lu *et al*., 2008).

Le modèle ARIMA est utilisé dans diverses études, pour la prévision des niveaux d'eau du lac Malawi (Makwinja *et al*., 2017), pour la prévision de la salinité de l'eau dans la baie d'Apalachicola en Floride (Sun et Koch, 2001), pour la prévision du dioxyde de soufre à Téhéran (Hassanzadeh *et al*., 2009). Chen *et al*. (2015) ont développé un modèle ARIMA pour prédire les concentrations quotidiennes de la Chl-*a* dans le lac Taihu en Chine et ont démontré son efficacité par rapport à une régression linéaire multivariée.

Le modèle saisonnier, autorégressif, à moyenne mobile intégrée (SARIMA) est composé du modèle ARIMA, incluant la composante saisonnière des données de la série chronologique.

SARIMA est très fréquemment utilisé pour les séries chronologiques mensuelles qui présentent un schéma saisonnier (Prista *et al*., 2011). La Chl-*a* est un paramètre connu pour être lié à la température (Tizro *et al*., 2014), qui a des caractéristiques saisonnières. Pour cette raison, un modèle SARIMA est mis en place pour traiter les caractéristiques des variations saisonnières, ce qui améliore la précision des prévisions.

Dans ce troisième chapitre, nous avons utilisé le modèle SARIMA pour prévoir les valeurs futures de la Chl-*a* dans la Lagune Nord de Tunis comme indicateur d'eutrophisation. Ce modèle a été appliqué avec une série temporelle d'environ trois décennies. Différents modèles SARIMA $(p, d, q)(P, D, Q)_S$ ont été mis en place.

Le modèle choisi, un SARIMA $(2,0,2)(2,0,2)_{12}$ avec les AIC et BIC (calcul de la perte d'information) les plus bas a été utilisé pour les prévisions.

La performance du modèle a été analysée en comparant avec les données relatives aux dernières observations, et en vérifiant les résultats des résidus. Le diagnostic des résidus a indiqué qu'ils sont non corrélés et relativement normalement distribués.

Les résultats des prévisions sont assez satisfaisants puisque la période de prévision semble reproduire relativement bien le contenu mensuel normal de la Chl-*a* dans la lagune.

Malgré le fait que la Lagune Nord de Tunis ait été classée comme zone humide d'importance internationale (site Ramsar), compte tenu de son histoire et de son importance, cet écosystème reste fragile.

Le modèle SARIMA appliqué aux données historiques de la Chl-*a* ou à tout autre paramètre clé, pourrait être un outil important pour fournir des informations avec une longueur d'avance qui orientent les interventions de prévention et de contrôle de l'écosystème.

Par rapport au lissage exponentiel (Moving Average) et au modèle de régression linéaire multivariée (MVLR), il a été déterminé que la technique de Box et Jenkins était supérieure (Jeong *et al*., 2014).

Par rapport au modèle MVLR et à la machine à vecteur de soutien (Support Vector Machine ou SVM), les techniques ANNs se sont révélées supérieures (Jeong *et al*., 2014).

Dans les chapitres précédents, le modèle SARIMA (Seasonal Autoregressive integrated moving average) et le modèle NARX (artificial neural network) ont donné de bons résultats dans la prévision du contenu de la Chl-*a*, dans la Lagune Nord de Tunis, comme indicateur

d'eutrophisation, en traitant les caractéristiques linéaires et non-linéaires de la série chronologique, respectivement.

Cependant, les problèmes réels ont généralement des caractéristiques à la fois linéaires et non linéaires, et des recherches antérieures ont tenté d'expliquer la variabilité des problèmes réels en combinant deux méthodes (Chen *et al*., 2007).

En effet, les lagunes côtières sont soumises à de fréquentes perturbations et fluctuations environnementales (évaporation, précipitations, ...), en plus de la relation étroite avec la température et la dépendance des saisons. C'est pourquoi le modèle hybride est devenu une pratique courante pour améliorer la précision des prévisions.

Le quatrième chapitre porte sur une récente technique de prédiction hybride pour la prévision du contenu de la Chl-*a* dans la Lagune Nord de Tunis en combinant le modèle SARIMA (approche linéaire) et le modèle NARX (approche non linéaire).

Ce chapitre a été réalisé en trois étapes : (1) établissement des données de concentrations de la Chl-*a* dans la Lagune Nord de Tunis ; (2) prévision de la variation mensuelle de la Chl-*a* en utilisant l'approche linéaire (modèle SARIMA) ; et (3) amélioration de la précision des prévisions en considérant la non-linéarité des résidus du modèle SARIMA en utilisant le modèle NARX.

Des recherches antérieures ont tenté d'améliorer la précision des prédictions des valeurs futures en combinant des approches linéaires et non linéaires. Qin *et al*. (2017) ont prédit la possibilité d'occurrence et les tendances futures de la marée rouge dans différentes zones côtières grâce à un modèle hybride qui combinait ARIMA et le Deep Belief Network (DBN).

Garcia Nieto *et al*. (2013) ont prédit la présence de cyanotoxines dans le réservoir de Trasona (nord de l'Espagne) grâce à un modèle de prévision hybride qui combinait la régression du vecteur de soutien avec des algorithmes génétiques. Shafaei *et al*. (2016) ont utilisé un modèle hybride qui combinait SARIMA et ANN pour la prévision des précipitations à Nahavand, en Iran.

En comparant le modèle hybride (SARIMA-NARX) réalisé dans ce quatrième chapitre, avec la technique SARIMA ou encore NARX mises en place dans les chapitres précédents, en utilisant des mesures mathématiques, le modèle hybride a montré de meilleures compétences ($R= 0.82$; $R^2=0.67$ ; $MSE= 0.24$).

SARIMA n'a pas obtenu de bons résultats dans la prévision des fluctuations de la Chl-*a*, en particulier des valeurs minimales et maximales, alors que le modèle hybride a amélioré la précision des prévisions sur ces points. Les performances du modèle hybride et du NARX sont relativement proches, avec une amélioration lors de l'utilisation de la méthode combinée.

Cette approche combinée pourrait permettre aux cadres et gestionnaires en charge de l'écosystème, d'estimer avec précision les concentrations mensuelles de la Chl-*a* comme indicateur d'eutrophisation dans la Lagune Nord de Tunis. Cette approche pourrait également être appliquée à d'autres écosystèmes naturels pour estimer tout paramètre clé (comme les nutriments, l'oxygène dissous, la salinité, le carbone, …) permettant d'évaluer le niveau d'eutrophisation.

Pour comparaison, dans chaque chapitre, nous avons simulé les mêmes approches sur les données historiques de la Station 5 (la plus affectée par l'eutrophisation).

Le RF et le NARX simulés dans le chapitre II, ont montré une petite baisse de performance lors de leurs simulations pour la Station 5, mais qui restent quand même assez performants (RF, $R^2$= 0.54; NARX, R=0.73). Ce qui peut être du à la grande hétérogénéité des données dans cette zone de la lagune. Le modèle RF a montré que le phosphore total et la salinité suivi de l'azote total sont les paramètres qui sont les plus liées à la variation de la Chl-*a* dans la Station 5.

Le SARIMA (1,1,0) $(0,1,0)_{12}$ simulé pour la Station 5 a montré la meilleure performance parmi les SARIMA testés.

Le modèle hybride développé pour la Station 5 a montré de meilleures résultats en comparaison avec le NARX et le SARIMA (1,1,0) $(0,1,0)_{12}$ simulé dans cette zone de la lagune (R= 0.78).

Les approches de modélisation décrites dans cette étude pour l'analyse et la prédiction du problème de l'eutrophisation dans la Lagune Nord de Tunis sont particulièrement utiles aux écologistes et aux environnementalistes, car elles leur permettront de prévoir les niveaux de pollution des eaux et de prendre à l'avance les mesures de précaution nécessaires.

Ces techniques ont la capacité d'identifier des paramètres importants pour permettre à la fois une surveillance physico-chimique sélective et une évaluation rapide et moins couteuse de la qualité de l'eau de la Lagune Nord de Tunis.

Les modèles développés peuvent être utilisés pour (1) estimer les concentrations de la Chl-*a* lorsque la valeur réelle n'est pas disponible et pour (2) simuler différents scénarios de la qualité de l'eau pour des valeurs extrêmes de paramètres clés et (3) réduire le coût et gagner du temps qu'en effectuant des travaux de terrain et des analyses en laboratoire que si nécessaire.

Même si la prévision à court terme est très intéressante pour prendre des décisions rapides et adéquates avec des résultats plus précis, elle représente aussi les limites de cette étude. Des travaux supplémentaires sur les algorithmes devraient être effectués pour essayer de faire des prévisions sur de longues périodes à l'avenir avec des performances plus précises, notamment avec la technique hybride.

En guise de recommandations pour la poursuite des recherches, de nombreux scénarios peuvent être réalisés à l'aide de ces modèles. Du plus optimiste au plus pessimiste, en mettant en place les valeurs extrêmes de la Chl-*a* ou de tout autre paramètre clé pouvant décrire le processus d'eutrophisation des écosystèmes naturels.

Lors de l'utilisation de la technique d'analyse de la variance (ANOVA) ou encore RF, les deux modèles ont considéré que la variation spatiale, semblait ne pas être significatif dans la variation de la Chl-*a* et des paramètres physico-chimiques dans la Lagune Nord de Tunis. Pour cette raison, on peut recommander également, de reconsidérer l'emplacement des cinq Stations dans la lagune. L'utilisation des techniques de ML peut aider à choisir d'autres zones dans l'écosystème étudié, qui seraient plus significatives dans le contrôle de la qualité de l'eau. D'après nos résultats (ANOVA et RF), il semblerait que moins de Stations peuvent être plus efficaces dans le contrôle de la qualité de l'eau, ce qui signifie gain de temps et d'argent.

Afin d'améliorer la précision des modèles, nous suggérons aussi d'ajouter plus de données, soit par simulation (en interpolant les données disponibles), soit idéalement en effectuant des mesures quotidiennes ou hebdomadaires, au moins pour les paramètres les plus importants (Chl-*a*, oxygène dissous, nutriments), peut-être pas à toutes les Stations mais au minimum à la Station 2 (dans la zone nord de la lagune) et la Station 5 (dans la zone sud de la lagune et la plus éloignée de la mer).

Pour finir, il convient de mentionner que ce type d'approche est extrêmement utilisé pour prévoir l'évolution de la plus grande crise sanitaire connue par le monde entier (la pandémie de coronavirus).

# Annexes