SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Earth and Environmental Sciences (DISAT)

PhD program in Chemical, Geological and Environmental Sciences.
Cycle XXXII

Curriculum in Environmental Sciences

# DATA DRIVEN APPROACH TO DEAL WITH DIFFERENT HYDROGEOLOGICAL ISSUES

Surname: Zanotti Name: Chiara

Registration number: 717614

Tutor: Dr Barbara Leoni

Co-tutor: Prof. Tullia Bonomi, Dr. Simone Sterlacchini
Supervisor: Dr. Marco Rotiroti

Coordinator: Prof. Marialuce Frezzotti

**ACADEMIC YEAR 2018/2019**

# Abstract

Due to the constantly growing interest toward environment protection, the amount of available data concerning environmental monitoring is increasing. As the size and complexity of environmental datasets continue to grow, there is a wide variety of possibility for implementation of data science in the environmental sciences field. The focus of the present PhD work is the resolution different hydrogeological issues by means of data science.

More specifically, the present PhD project aims at identifying and applying data-driven techniques suitable for hydrogeological datasets, based on the structure of the problem and the available data and on site-specific conditions. In the scope of this PhD work two main hydrological problems were addressed parallelly, concerning the two main aspects of groundwater resource management: a) groundwater quality and b) groundwater quantity. Each task was tackled in two successive phases. The first phase consisted in an exploratory analysis of the available data, aimed at reaching a better understanding of the system, the problem and the available information. The second phase involved the application of specific data driven techniques to investigate their effectiveness in the hydrogeological field.

The groundwater quality data analysis involves the application of multivariate techniques, normally used for the source apportionment, to a dataset concerning chemical data of surface water and groundwater aiming at determining their effectiveness in identifying the phenomena that contribute to the concentration of several compounds in a sample. In this task Factor Analysis, Cluster Analysis and Positive Matrix Factorization were implemented.

As regards the groundwater quantity, the analysis of groundwater level time series uses models able to reconstruct historical data and applicable to forecast scenarios; in this task autocorrelation, partial autocorrelation and impulse response were used and linear and nonlinear neural networks models were developed.

This work highlighted that data driven techniques can be considered useful tool to support groundwater resource management.

# Graphical abstract

## Chapter 1 - INTRODUCTION

Physically based models

Groundwater (GW) management

Simulate 3D flow path

Quantifying exchanges and water budget

high resolution GW level forecasting

Multivariate GW quality data management

Data driven techniques

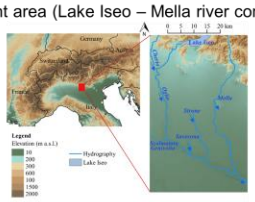
## Chapter 2 - GROUNDWATER QUALITY
## Data mining of GW quality dataset

### Main aims

- To identify the main hydrochemical features of either groundwater and surface water (lake, rivers and springs) and, possibly, the processes that influence (or govern) them
- To understand the chemical relations (if any) between groundwater and surface water

**Study area:**

Oglio catchment area (Lake Iseo – Mella river confluence)

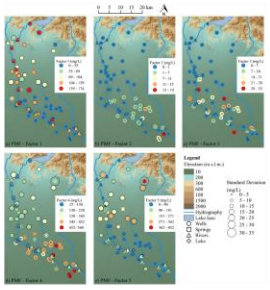


### Phase 1 – Exploratory analysis

**Multivariate statistical analysis supporting the hydrochemical characterization of groundwater and surface water: a case study in northern Italy.** Rend Online Soc Geol It, 47:90-96.

### Phase 2 – PMF model

**Groundwater and surface water quality characterization through positive matrix factorization combined with GIS approach.** WATER RESEARCH, 2019 (159), 122-134.



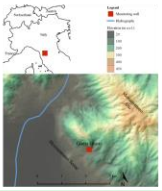
## Chapter 3 - GROUNDWATER QUANTITY
## GW level forecasting

### Main aims

- To study the relationship between GW level and recharge from precipitation
- Short and long term groundwater level forecasting as a tool to support a more efficient and sustainable groundwater resource management

**Study area:**

Monsummano Terme (PT), Tuscany

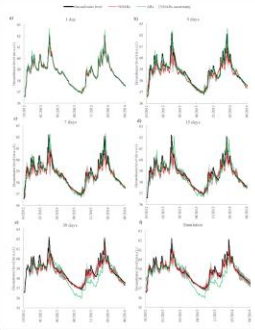


### Phase 1 – Exploratory analysis

**Groundwater level forecasting using linear time series modeling: the case study of the thermal aquifer system of Monsummano Terme (central Italy).** Rend. Online Soc Geol It 47: 153–160.

### Phase 2 – Neural Networks

**Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system.** JOURNAL OF HYDROLOGY, 2019 (578),124015



## Chapter 4 - CONCLUSIONS

In a landscape of increasing availability of environmental data, data science can actually become a useful instrument to enhance the understanding of groundwater systems and to promote a more sustainable and efficient use of the groundwater resource

*Chiara Zanotti – XXXII Ciclo*

# Table of contents

# Acknowledgements

I would like to express my sincere gratitude to Prof. Tullia Bonomi, for foreseeing the potential of the unconventional approach of this research project, for creating this line of work, and for giving me the possibility and the freedom to learn, try, fail and improve, and for trusting me during the whole process.

I would also like to thank the experts who supervised this thesis: Dr. Barbara Leoni, Dr. Simone Sterlacchini and Dr. Marco Rotiroti, for their precious contribution to these research outputs. Furthermore, I would like to thank Dr. Letizia Fumagalli for her continuous support and cheering.

I would like to thank Prof. Andre Prévôt, and Dr. Francesco Canonaco from the Paul Scherrer Institute for sharing their knowledge and tools for the PMF application, and Dr. Giulia Stefenelli for her technical support, warm hospitality and everlasting friendship.

I would like to thank the laboratories of DISAT, in particular dott. Valentina Soler and dott. Maria Tringali for providing hydrochemical analysis. I would also like to thank dott. Giacomo Cappellini for his practical support about machine learning. Furthermore, dott. Martina Patelli and dott. Veronica Nava for their contribution to the Cariplo project and for the enjoyable time we spent together.

My sincere thanks also go to Dr. Sara Taviani for her genuine interest in my work and my life, and for the meaningful time we shared.

A special thanks go to my everyday office mates for all the precious moments we shared. Thanks to Dr. Alberto Stefania, for all the laughs, the sorrow and the whole emotional roller coaster we went through together and for our unique friendship, on the right. Thanks to dott. Davide Sartirana, for his contagious irony and auto-irony and his thoughtful presence.

I would also like to thank my family, especially my sister for her always wise suggestions.

Last but definitively not least, I would like to thank my husband, for his infinite support and pushing, for believing in me even when I did not. Thank you.

# List of Abbreviations

| | |
|---|---|
| ACF | Autocorrelation function |
| AIC | Akaike Information Criterion |
| AR | Autoregressive model |
| ARx | Autoregressive model with exogenous input |
| CA | Cluster Analysis |
| FA | Factor Analysis |
| GW | Groundwater |
| MSE | Mean Squared Error |
| nAIC | standardized Akaike Information Criterion |
| NNARx | Neural Network – ARx combined models |
| PACF | Partial Autocorrellation Function |
| PMF | Positive Matrix Factorization |
| RMSE | Root Mean Squared Error |
| SSE | Sum of Squared Errors |

# Chapter 1
# **Introduction**

The concept of *data*, as defined by Gould (1971) is: "*DATA: a representation of facts or ideas in a formalised manner capable of being communicated or manipulated by some process"*. Consequentially, Naur (1974) defines *data science* as *" the science of dealing with data, while the relation of data to what they represent is delegated to other fields and sciences".*

Data science is a wide concept including data gathering from multiple sources, cleansing, preparation, and different kind of analysis (e.g. machine learning, predictive analytics etc.) to extract critical information from the collected data sets.

In the last decades, data science methods obtained an increased popularity and they seems to be of interest for many organizations, including those working on the management of environmental system. Data science is widely used in a variety of fields including business, bioinformatics, marketing and sentiment analysis, energy, smart cities etc. and several sectors are facing a significant transformation as a result of the diffusion of data-driven innovation. A similar transformation is underway among several scientific disciplines, including environmental sciences, to investigate the benefits that can be realized through use of appropriate data science approaches (Gibert et al., 2018).

The complexity and the size of environmental datasets continue to grow, and there is a variety of opportunity for applications of data science in the environmental sciences field (Gibert et al., 2008; Kanevski et al., 2008).

With the evolution of environmental sciences, environmental datasets are growing in size, complexity and resolution. Due to the increasing attention toward environment protection by both policymakers and stakeholders, monitoring environmental phenomena is becoming progressively more frequent. This leads to an increasing amount of environmental data stored but often not really exploited. In the hydrology field, protection of the water resources happens through a constant and spatially detailed monitoring of the quantitative and qualitative status of the resource. This means that the amount of available data representing the phenomena related to the water resource is constantly increasing. In most of cases those data are only used to evaluate the current situation with respect to the current regulations, but their value could be more exploited by identifying the information hidden among them.

The process of extracting information concealed within the mass of large databases is called *data mining*. Data mining is a branch of the family of the machine learning techniques which focuses on the so called unsupervised learning, consisting in finding groups or cluster of data or variables among a dataset. The success of data mining is mainly connected to the possibility that there is valuable information in the data that one already has and that it could be teased out (Hand, 2006).

The possibility of exploring the information contained in datasets characterized by several variables allows for a deeper understanding of the sources affecting the groundwater resources quality, and understanding the natural and anthropogenic pressures active on a system is the basis for the development of proper water quality monitoring, assessment and preservation.

Furthermore, the increasing amount of data resulting from the enlarged attention toward water resources protection, becomes even more valuable if considered as the raw material for the developing of forecasting tools. A second brunch of the data science, which focuses on the supervised learning, aims at building mathematical model from a set of data that contains both the inputs and the desired outputs. The increasing availability of on-line sensors allows to produce and store significant amount of data related to time oscillation of natural and anthropogenic phenomena. Machine learning can extract the information contained in the data and use it for training mathematical model able to reproduce those data and make future projections.

The possibility of developing data driven groundwater level forecasting tools becomes particularly appealing if considering that they do not require prior information about the structure and the composition of the aquifer as opposite to numerical flow models (Maier and Dandy, 2000).

Indeed, numerical flow models can support groundwater management by providing a wide variety of information such as flow directions and water budget (Sun, 2013) but they also require a lot of information for a proper calibration (Burrows and Doherty, 2016, 2015). Contrarily, data driven models, offer a more agile solution when only forecasts on single locations are needed and the information about system structure is limited. Therefore, data driven models can become useful tools, leading to more sustainable water supply management solutions, allowing decision makers to activate plans, apply strategies and take proactive measures towards groundwater resource protection when groundwater levels is predicted to be decreased below the normal levels (Kouziokas et al., 2018; Wunsch et al., 2018).

The main aim of the present PhD work is to explore the potential of data science in the hydrogeology field, by testing the use of data driven techniques for solving different hydrogeological issues. The PhD project was focused on two main aspects of the hydrogeology field, which are groundwater quality and groundwater quantity. The structure of the present PhD work was consequentially divided into two parallel tasks:

- groundwater quality – multivariate analysis of chemical variables in water samples
- groundwater quantity – groundwater level forecasting

Both tasks were developed in two phases: the first phase involves the exploratory analysis of the data, while the second one the implementation of the investigated models, the interpretation of the results and of their uncertainties.

Chapters 2 and 3 illustrate the two tasks of this PhD, i.e. groundwater quality and groundwater quantity. Each chapter is composed by an introduction, followed by the description of the two phases of each task.

## 1.1 Groundwater quality – multivariate analysis of chemical variables in water samples.

The main aim of this task is to perform data mining of a groundwater quality data set by implementing multivariate statistical analysis as a tool. This part of the PhD work is focused on the spatial variability of chemical variables.

Multivariate analysis is a branch of the statistics which deals with observations characterized my many variables. The main goal is to study how the variables are related to each another, and how they work in combination to distinguish between the different cases in which the observations were made. All physical, biological and natural processes are fundamentally multivariate in nature; the challenge of multivariate analysis is to comprehend the process in a multivariate way, considering the variables as connected and understanding their relationships, as opposed to a bunch of univariate processes, i.e. single variables at a time, isolated from one another (Greenacre and Primicerio, 2014).

The data driven techniques used for this task are Cluster Analysis (CA), Factor Analysis (FA), and Positive Matrix Factorization (PMF). CA aims at identifying groups of samples with the same chemical characteristics, while FA and PMF aim at identifying latent factors governing groups of variables. PMF (Paatero and Tapper, 1994) is a bilinear model in which the fundamental problem is to solve the identity and contribution of several sources on several samples.

In the last decades, attention toward multivariate statistical analysis as a tool to analyse water chemical data increased. Several authors applied FA or CA to explore groundwater and surface water chemical data and identify natural phenomena (Blake et al., 2016; Koh et al., 2016; Shrestha et al., 2016) as well as anthropogenic impacts affecting water quality (Alberti et al., 2016; Devic et al., 2014; Gu et al., 2018; Phung et al., 2015). PMF on the other hand, was specifically designed for air pollution data, and only in a few cases it was adopted for data related to other environmental matrices. The main aim of this task is to test the effectiveness of PMF as a tool to perform data analysis of water quality data and compare it with a more widely used FA.

## 1.2 Groundwater quantity – Groundwater level forecasting

The main aim of this task is to analyze groundwater level time variability through time series analysis and to develop a forecasting tool. This part of the PhD work was focused on the time variability of the groundwater level in a single monitoring station.

In descriptive statistics, a time series is defined as a set of random variables with respect to time and expresses the dynamics of a certain phenomenon over time. The time series are studied both to understand a phenomenon, identifying trend components, cyclicality, seasonality and / or randomness, and to forecast its future trend. The peculiarity of time series analysis lays in possibility of taking into account for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. Here, groundwater level time series are analysed both, in a univariate and multivariate manner. Autocorrelation (ACF) and Partial Autocorrelation Function (PACF) were used as univariate analysis to determine the "memory effect" of the groundwater level data, and the univariate Autoregressive linear model (AR) was implemented. Cross correlation function and Impulse Response (IR) were implemented as bivariate analysis to investigate over time the relationship between groundwater level and precipitation data. Furthermore, forecasting models were developed, particularly: Autoregressive model with exogenous input (ARx) and Neural Network – ARx combined models (NNARx).

## 1.3 References

Alberti, L., Azzellino, A., Colombo, L., Lombi, S., 2016. Cluster Analysis to identify Tetrachloroethylene pollution hotspots for transport numerical model implementation in urban functional area of Milan, Italy, in: 16th International Multidisciplinary Scientific Conference SGEM2016, Book 1. https://doi.org/10.5593/SGEM2016/B11/S02.091

Blake, S., Henry, T., Murray, J., Flood, R., Muller, M.R., Jones, A.G., Rath, V., 2016. Applied Geochemistry Compositional multivariate statistical analysis of thermal groundwater provenance : A hydrogeochemical case study from Ireland. Appl. Geochemistry 75, 171–188. https://doi.org/10.1016/j.apgeochem.2016.05.008

Burrows, W., Doherty, J., 2016. Gradient-based model calibration with proxy-model assistance. J. Hydrol. 533, 114–127.

Burrows, W., Doherty, J., 2015. Efficient calibration/uncertainty analysis using paired complex/surrogate models. Groundwater 53, 531–541.

Devic, G., Djordjevic, D., Sakan, S., 2014. Natural and anthropogenic factors affecting the groundwater quality in Serbia. Sci. Total Environ. 468–469, 933–942. https://doi.org/10.1016/j.scitotenv.2013.09.011

Gibert, K., Horsburgh, J.S., Athanasiadis, I.N., Holmes, G., 2018. Environmental Modelling & Software Environmental Data Science. Environ. Model. Softw. 106, 4–12. https://doi.org/10.1016/j.envsoft.2018.04.005

Gibert, K., Spate, J., Sànchez-Marrè, M., Athanasiadis, I.N., Comas, J., 2008. Chapter twelve data mining for environmental systems. Dev. Integr. Environ. Assess. 3, 205–228.

Gould, I.H., 1971. IFIP Guide to concepts and terms in data processing. Amsterdam: North-Holland Publishing Company.

Greenacre, M., Primicerio, R., 2014. Multivariate analysis of ecological data. Fundacion BBVA.

Gu, H., Ma, F., Guo, J., Zhao, H., Lu, R., Liu, G., 2018. A Spatial Mixing Model to Assess Groundwater Dynamics Affected by Mining in a Coastal Fractured Aquifer, China. Mine Water Environ. 37, 405–420. https://doi.org/10.1007/s10230-017-0505-x

Hand, D.J., 2006. Data Mining. Encycl. Environmetrics 2.

Kanevski, M., Pozdnukhov, A., Timonin, V., 2008. Machine learning algorithms for geospatial data. Applications and software tools.

Koh, D.C., Chae, G.T., Ryu, J.S., Lee, S.G., Ko, K.S., 2016. Occurrence and mobility of major and trace elements in groundwater from pristine volcanic aquifers in Jeju Island, Korea. Appl. Geochemistry 65, 87–102. https://doi.org/10.1016/j.apgeochem.2015.11.004

Kouziokas, G.N., Chatzigeorgiou, A., Perakis, K., 2018. Multilayer Feed Forward Models in Groundwater Level Forecasting Using Meteorological Data in Public Management. Water Resour. Manag. 5041–5052.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables : a review of modelling issues and applications 15, 101–124.

Naur, P., 1974. Concise survey of computer methods. Petrocelli Books.

Paatero, P., Tapper, U., 1994. Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. Environmetrics 5, 111–126.

Phung, D., Huang, C., Rutherford, S., Dwirahmadi, F., Chu, C., Wang, X., Nguyen, M., Nguyen, N.H., Do, C.M., Nguyen, T.H., Dinh, T.A.D., 2015. Temporal and spatial assessment of river surface water quality using multivariate statistical techniques: a study in Can Tho City, a Mekong Delta area, Vietnam. Environ. Monit. Assess. 187. https://doi.org/10.1007/s10661-015-4474-x

Shrestha, S.M., Rijal, K., Pokhrel, M.R., 2016. Assessment of Heavy Metals in Deep Groundwater Resources of the Kathmandu. J. Environ. Prot. (Irvine,. Calif). 7, 516–531.

Sun, N.-Z., 2013. Inverse problems in groundwater modeling. Springer Science & Business Media.

Wunsch, A., Liesch, T., Broda, S., 2018. Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX). J. Hydrol. 567, 743–758. https://doi.org/10.1016/j.jhydrol.2018.01.045

# Chapter 2
# Groundwater quality

## 2.1 Introduction

Groundwater constitutes the first source of drinking and irrigation water in many countries worldwide. This phenomenon makes groundwater quality strictly connected with human health (Barzegar et al., 2018; Ritter et al., 2002), leading several regions worldwide to promulgate national and supranational regulations aimed at protecting groundwater resources.

EU water policy aims at ensuring that a sufficient quantity of good quality water could be available to fulfil both the ecosystems and people's need. The Water Framework Directive, promulgated in 2000 (European Community, 2000; European Union, 2006) determined a framework for the improvement of the quality of EU water resources involving assessment, management and protection actions.

Generally, the costs and complexity of water treatment plants for potabilization and remediation actions is proportional to pollution of the original water source.

Groundwater quality can be affected by a variety of factors involving anthropogenic and natural phenomena (Tziritis et al., 2016). Anthropogenic impact is strictly related to the landuse (agricultural, urban or industrial) and can determine both diffuse or punctual contaminations of a wide variety of compounds. A wide range of natural phenomena largely contribute in determining the status of the groundwater resources. First of all, the structure and composition of the subsoil structure, which directly interacts with the water and determines whether the resource can be reached by fresh recharge (Robinson and Ayotte,

2006). Secondly, groundwater interacts with surface water bodies such as rivers and lakes. Exchanges can happen in both ways which means that groundwater and surface water qualities are strictly connected (Feinstein, 2012; Winter et al., 1998).

Therefore, to assess groundwater quality, it becomes imperative to firstly identify which are the natural and anthropogenic phenomena that influence the water status. Understanding and quantifying the pressures active on the water resources are the first steps for guaranteeing a proper water resource monitoring and management. In this scope, especially in certain areas, understanding which are the relationship between groundwater and surface water can be crucial for ensuring the maintenance of a good status of the water resources (Kalbus et al., 2006; Sophocleous, 2002).

This task addresses the problem of managing, visualizing and understanding big datasets of chemical data concerning groundwater and surface water. Chemical data collected during field surveys are the numerical expression of several natural and anthropic processes that take place within a given environmental system. To manage those data with a data-driven approach allows to extract from them as much information as possible about the processes that led to them, identifying those processes (natural phenomena, anthropic exploitation, land use impacts etc.) and quantifying their contributes.

The study area is part of the Oglio River basin, in the Po plain (Northern Italy), between the outflow from Lake Iseo and the confluence into Mella River, and covers ~1900 $km^2$. The study area presents different geological characteristics from north to south: the higher plain in the northern part hosts a mono-layer aquifer mainly composed of sands and gravels, while the lower plain in the southern region hosts a multi-layer aquifer constituted by a vertical alternation of sands with silty clays. At the transition zone between the two regions numerous natural groundwater outflows are present determining the so called "spring belt".

The whole study area, strongly impacted by agricultural landuse, is characterized by strong relationships between groundwater and several surface water bodies such as rivers, channels, springs and Lake Iseo.

This work was carried in the scope of the project "Lake, stream and groundwater modelling to manage water quantity and quality in the system of Lake Iseo – Oglio River" supported by

Fondazione Cariplo (grant number 2014-1282) brought off between 2015 and 2018. Two of the main aims of the project were: a) to identify the main hydrochemical features of either groundwater and surface water (lake, river and springs) and the processes that influence and govern them, and b) to understand the relations between the chemistry of groundwater and surface water. In the scope of the project, samples were gathered from wells, springs, rivers and Lake Iseo and analysed for several chemical compounds.

In this task of the present PhD work data analysis of the resulting dataset was carried out, in order to investigate whether data driven techniques could be of use in characterizing groundwater and surface water of the study area. This task was performed in two successive phases:

1) The first phase, entitled "Multivariate statistical analysis supporting the hydrochemical characterization of groundwater and surface water: a case study in northern Italy" consists in the exploratory analysis of the dataset resulting from a first field survey, conducted in Fall 2015. Here, FA and CA were applied for a first identification of the major phenomena represented by the data. The results of this exploratory analysis were used to define a reduced monitoring network that was used for the successive field sampling campaign of 2016-2017 by avoiding redundant sampling points, and wells affected by phenomena that were not considered in the main focus of the project.

2) The second phase, entitled "Groundwater and surface water quality characterization through positive matrix factorization combined with GIS approach" consists in the analysis of the data collected during four seasonal field surveys conducted in the study area from February 2016 until March 2017, representing a whole hydrological year. Here, the use of PMF was investigated as an alternative to FA (Paatero and Tapper, 1994). PMF is a multivariate analysis aimed at source identification and apportionment. It was specifically designed to cope with environmental data and to manage their uncertainty (Paatero, 2000). In the last decades PMF was widely used in the air pollution field (Visser et al., 2009) while recently a few studies demonstrated that PMF can be successfully applied to datasets concerning different environmental matrices, e.g. soil and lake sediments, to reach a more realistic representation of the sources affecting different systems (Comero et al., 2009). Here, the effectiveness of PMF as a

tool to perform groundwater and surface water characterization is investigated, and its results are compared with those obtained with the more widely used FA.

**References**

Barzegar, R., Asghari, A., Jan, M., 2018. Assessing the potential origins and human health risks of trace elements in groundwater : A case study in the Khoy plain , Iran. Environ. Geochem. Health 9. https://doi.org/10.1007/s10653-018-0194-9

Comero, S., Capitani, L., Gawlik, B.M., 2009. Positive Matrix Factorisation (PMF) An introduction to the chemometric evaluation of environmental monitoring data using PMF. https://doi.org/10.2788/2497

European Community (2000) - Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Official Journal of the European Parliament, L327(September 1996), 1–82. https://doi.org/10.1039/ap9842100196

European Union (2006) - Directive 2006/118/EC of the European Parliament and of the council of 12 December 2006 on the protection of groundwater against pollution and deterioration. Official Journal of the European Union, 19(L372), 19–31. https://doi.org/http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32006L0118

Feinstein, D., 2012. Since "Groundwater and Surface Water--A Single Resource": some US Geological Survey advances in modeling groundwater/surface-water interactions. Acque Sotter. J. Groundw. 1.

Kalbus, E., Reinstorf, F., Schirmer, M., 2006. Measuring methods for groundwater and surface water interactions: a review. Hydrol. Earth Syst. Sci. Discuss. 10, 873–887.

Paatero, P., 2000. User's guide for the multilinear engine program "ME2" for fitting multilinear and quasi-multilinear models. Univ. Helsinki, Finl.

Paatero, P., Tapper, U., 1994. Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. Environmetrics 5, 111–126.

Ritter, L., Solomon, K., Sibley, P., Hall, K., Keen, P., Mattu, G., Linton, B., 2002. Sources, pathways, and relative risks of contaminants in surface water and groundwater: a perspective prepared for the Walkerton inquiry. J. Toxicol. Environ. Heal. Part A 65, 1–142.

Robinson, G.R., Ayotte, J.D., 2006. The influence of geology and land use on arsenic in stream sediments and ground waters in New England, USA. Appl. Geochemistry 21, 1482–1497.

Sophocleous, M., 2002. Interactions between groundwater and surface water: the state of the science. Hydrogeol. J. 10, 52–67. https://doi.org/10.1007/s10040-001-0170-8

Tziritis, E., Skordas, K., Kelepertsis, A., 2016. The use of hydrogeochemical analyses and multivariate statistics for the characterization of groundwater resources in a complex aquifer system . A case study in Amyros River basin , Thessaly , central Greece. Environ. Earth Sci. 75, 1–11. https://doi.org/10.1007/s12665-015-5204-y

Visser, A., Dubus, I., Broers, H.P., Brouyère, S., Korcz, M., Orban, P., Goderniaux, P., Batlle-Aguilar, J., Surdyk, N., Amraoui, N., Job, H., Pinault, J.L., Bierkens, M., 2009. Comparison of methods for the detection and extrapolation of trends in groundwater quality. J. Environ. Monit. 11, 2030–2043. https://doi.org/10.1039/b905926a

Winter, T.C., Harvey, J.W., Franke, O.L., Alley, W.M., 1998. Ground water and surface water; a single resource, Circular. https://doi.org/10.3133/cir1139

## 2.2 Multivariate statistical analysis supporting the hydrochemical characterization of groundwater and surface water: a case study in northern Italy

Marco Rotiroti *(1), Chiara Zanotti (1), Letizia Fumagalli (1), Sara Taviani (1), Gennaro A. Stefania (1), Martina Patelli (1), Veronica Nava (1), Valentina Soler (1), Elisa Sacchi (2) and Barbara Leoni (1).

(1) Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan, Italy
(2) Department of Earth and Environmental Sciences, University of Pavia, Pavia, Italy
*corresponding author; address: Piazza della Scienza 1, 20126 Milano, Italy; Tel: +39 02 64482882; Fax: +39 02 64482895; email: marco.rotiroti@unimib.it

**Rend. Online Soc. Geol. It., Vol. 47 (2019), pp. 90-96**
**Doi:** https://doi.org/10.3301/ROL.2019.17

### 2.2.1 Abstract

Multivariate statistical analysis is a useful method for supporting the interpretation of experimental data, particularly in the case of large datasets. In the present study, cluster analysis and factor analysis are used to support the hydrochemical characterization of groundwater and surface water in an area located in the Oglio River basin (N Italy).

During a field survey performed in Fall 2015, 58 groundwater, 20 river (Oglio River and its main tributaries), 1 Lake Iseo and 7 spring samples were collected for chemical analysis.

Results of multivariate statistical analysis led to the identification of the following 5 main clusters which characterize the hydrological system: (1) higher plain groundwater and springs, characterized by an oxidized *hydrofacies* with higher $NO_3$, (2) lower plain groundwater, characterized by a reduced *hydrofacies* with higher As, Fe and Mn, (3) Oglio River, (4) Oglio River tributaries and (5) outliers. This characterization will bear the improvement of the hydrogeological conceptual model of the area, also oriented to groundwater/surface water interactions, that, in turn, will support the numerical flow modeling of the system.

KEY WORDS: Cluster Analysis; Factor Analysis; Nitrate; Arsenic; Oglio River.

## 2.2.2 Introduction

Large experimental datasets require proper techniques and tools to be managed, elaborated and interpreted. Multivariate statistical analysis can be used for supporting the interpretation of environmental data (Facchinelli et al. 2001, Angelone et al., 2009, Re et al. 2014, Palmucci et al., 2016). The present study concerns the application of cluster analysis (CA) and factor analysis (FA) in order to support the hydrochemical characterization of groundwater and surface water in an area located in the Po Plain (N Italy), highly impacted by both human activities related to agriculture and natural contaminations. The former is mainly related to nitrate pollution, a well-known environmental problem affecting large areas of the Po Plain (Mantovi et al., 2006; Sacchi et al., 2013; Bonfanti et al., 2016; Lasagna et al., 2016) and other agricultural lands of Italy (Capri et al., 2009), whereas the latter mainly concerns arsenic pollution, another well-known environmental problem affecting the Po Plain aquifers (Carraro et al., 2015; Rotiroti et al., 2015a; Rotiroti et al., 2015b; Dalla Libera et al., 2017; Rotiroti et al., 2017).

The hydrochemical characterization of a hydrological system is a fundamental step needed in all hydrological studies concerning the evaluation of water resources quality. However, the different water bodies that compose a hydrological system (e.g. rivers, lakes, groundwater, etc.) are usually characterized individually avoiding a holistic approach. A relevant aspect of the present work is that the several components of the hydrological system of the study area (i.e. groundwater, Lake Iseo, Oglio River, its tributaries and springs) are jointly characterized.

## 2.2.1 Materials and methods

### *2.2.1.1 Study area*

The study area is located in the Oglio River basin, between the outflow from Lake Iseo and the confluence into Mella River, and covers ~1900 km$^2$ (Fig. 1a and 1b). The northern part of the study area (higher plain) hosts a mono-layer unconfined aquifer mainly composed of sands and gravels, whereas the southern part of the area (lower plain) hosts a multi-layer aquifer constituted by a vertical alternation of sands with silty clays (Fig. 1c); this multi-layer system can be subdivided into shallow (depth < 40 m b.g.s.), intermediate (between 40 and 100 m b.g.s.) and deep (> 100 m b.g.s.) aquifer sub-units. These aquifers are confined with some exceptions for the shallow aquifer that can become locally semi-confined or unconfined due

to local factors (e.g. local absence of shallow silty-clay confining layers). The transition between higher and lower plain is marked by the so called "spring belt", an area where numerous natural groundwater outflows are present. These hydrogeological features are common in the whole Po Plain (Bonomi, 2009; Mastrocicco et al., 2014; Perego et al., 2014). Previous studies (Buizza, 2012; Delconte et al., 2014; Rotiroti et al., 2016) revealed that the Oglio River is losing from the outflow from Lake Iseo to approximately 20-30 km downstream, then it becomes gaining up to the confluence into Mella River. In the unconfined aquifer of the higher plain, the groundwater table has a depth of ~50 m b.g.s. in the northern part, progressively decreasing to a few meters b.g.s. moving south towards the spring belt and the lower plain; groundwater in the higher plain mainly flows from N to S with a slight shift towards the Oglio River where it is gaining (Fig. 1b; Éupolis Lombardia, 2015). In the lower plain, the groundwater table depth approaches ground level; in the shallow aquifer, the groundwater flow direction is influenced by the gaining behaviour of Oglio River and its tributaries (Fig. 1b) whereas in the intermediate and deep aquifers groundwater mainly flows from NW to SE, that is the regional groundwater flow direction (Éupolis Lombardia, 2015). The Lake Iseo tends to fully circulate only irregularly, during harsh and windy winters (i.e. only two deep water mixing during 2005 and 2006 in the last 35 years; Leoni et al., 2014; Valerio et al., 2015) therefore, only the shallower lake water can be considered as water input to the Oglio River.

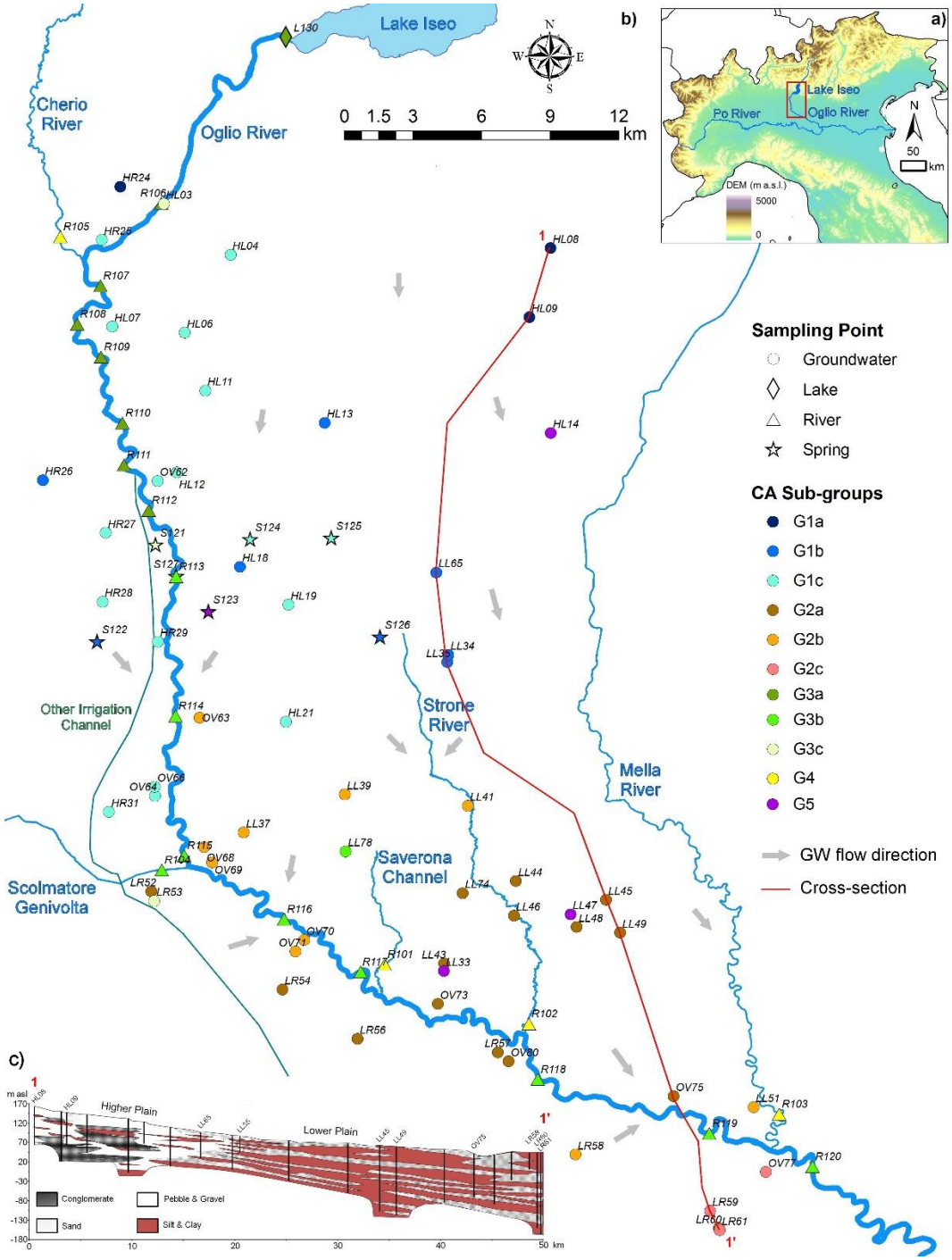



**Figure 1. a) Location of study area. b) Location and type of sampling points; colors represents the different groups and sub-groups resulting from cluster analysis (see the text for details); groundwater flow directions were deduced from the potentiometric map made by Éupolis Lombardia (2015), groundwater flow directions within the lower plain are referred to the shallow aquifer. c) Lithological cross-section outlining the aquifer systems of higher and lower plains.**

#### *2.2.1.2 Hydrochemical data*

During a field survey performed from October 26th to November 16th 2015, 58 groundwater, 20 river (Oglio River and its main tributaries), 1 Lake Iseo and 7 spring samples were collected for chemical and isotopic analysis. In the higher plain, the screens of the wells tapping the mono-layer unconfined aquifer are entirely positioned between 20 and 90 m b.g.s. In the lower plain, the sampled wells selectively tap the different overlaying aquifers that form the multi-layer system covering the whole 20-190 m b.g.s. depth interval. In all the 86 water samples, physico-chemical parameters (water temperature, pH, electrical conductivity (EC) and dissolved oxygen (DO)), major ions (alkalinity, Ca, Mg, Na, K, Cl, $SO_4$, $NO_3$ and $NH_4$) trace elements (As, Fe and Mn) and water isotopes ($\delta^2H$ and $\delta^{18}O$) were measured.

Before sampling, water wells were purged until physico-chemical parameters were constant,

generally after 2−3 well volumes were removed. Samples were filtered through 0.2 μm filters in the field, those for As, Fe and Mn analysis were acidified with nitric acid; after collection, samples were stored in a portable fridge at 4°C. Water temperature, EC, pH, and DO were measured in the field using the WTW Multi 3430 meter in a flow cell. Alkalinity was analysed by HCl tritation within 24 hours from the sampling. Major ions were analysed by ion chromatography. Ammonium was analysed by spectrophotometry with Nessler's reagent within 24 hours from the sampling. Iron and manganese were analysed by Inductively Coupled Plasma - Optical Emission Spectroscopy (ICP-OES) whereas arsenic was analysed by Graphite Furnace Atomic Absorption Spectrometry (GFAAS). Water isotopes were analysed by wavelength-scanned cavity ring-down spectroscopy (WS-CRDS).

#### *2.2.1.3 Multivariate statistical analysis*

The cluster analysis (CA) was performed on total 18 variables and 86 samples; data were auto-scaled (i.e. mean value = 0 and variance = 1). The Ward hierarchical method (Ward, 1963), based on the squared Euclidean distance as a measure of similarity between samples, was used. This method is used to group commonly measured water quality parameters; the resulting clusters indicate different types of water with particular features (Gibrilla et al., 2011).

The factor analysis (FA) was conducted using 18 variables and 82 samples, since 4 samples (LR59, LR61, OV77 and S123) were excluded from this analysis. These exclusions were due to (a) disproportionately high As values for samples LR59, LR61, OV77 and (b) a particular chemical composition of sample S123 allowing it to be considered as an outlier. Results of a preliminary FA including also these 4 outliers (not shown) evidenced mainly the difference between these 4 samples and all the other samples, hiding most of the information contained in the data. The eigenvalues for each factor were calculated with the auto-scaling method. The Kaiser criterion (Kaiser, 1958) was used to determine the significant factors. This method concerns the selection of those factors for which the eigenvalues are higher or equal to 1.

The software IBM SPSS® was used for performing both CA and FA.

Results of multivariate statistical analysis were combined with the geomorphological and hydrogeological knowledge of the study area in order to give a hydrological explanation of each data cluster.

### 2.2.2 Results and discussions

Results of CA are showed by the dendrogram in Fig. 2. Locations of sampling points for each cluster are reported in Fig. 1b. The CA grouped the sampling points into 16 clusters within the first grouping level, however, in order to give a hydrological and hydrogeological explanation of each data cluster, data were firstly associated to 5 main groups, then up to 3 sub-groups were identified for each main group. The following 5 main groups of data were identified (Fig. 2): (G1) higher plain groundwater and springs, (G2) lower plain groundwater, (G3) Oglio River and Lake Iseo, (G4) Oglio River tributaries and (G5) outliers. Mean values of measured chemical parameters for each group and sub-group are showed in Tab. 1.

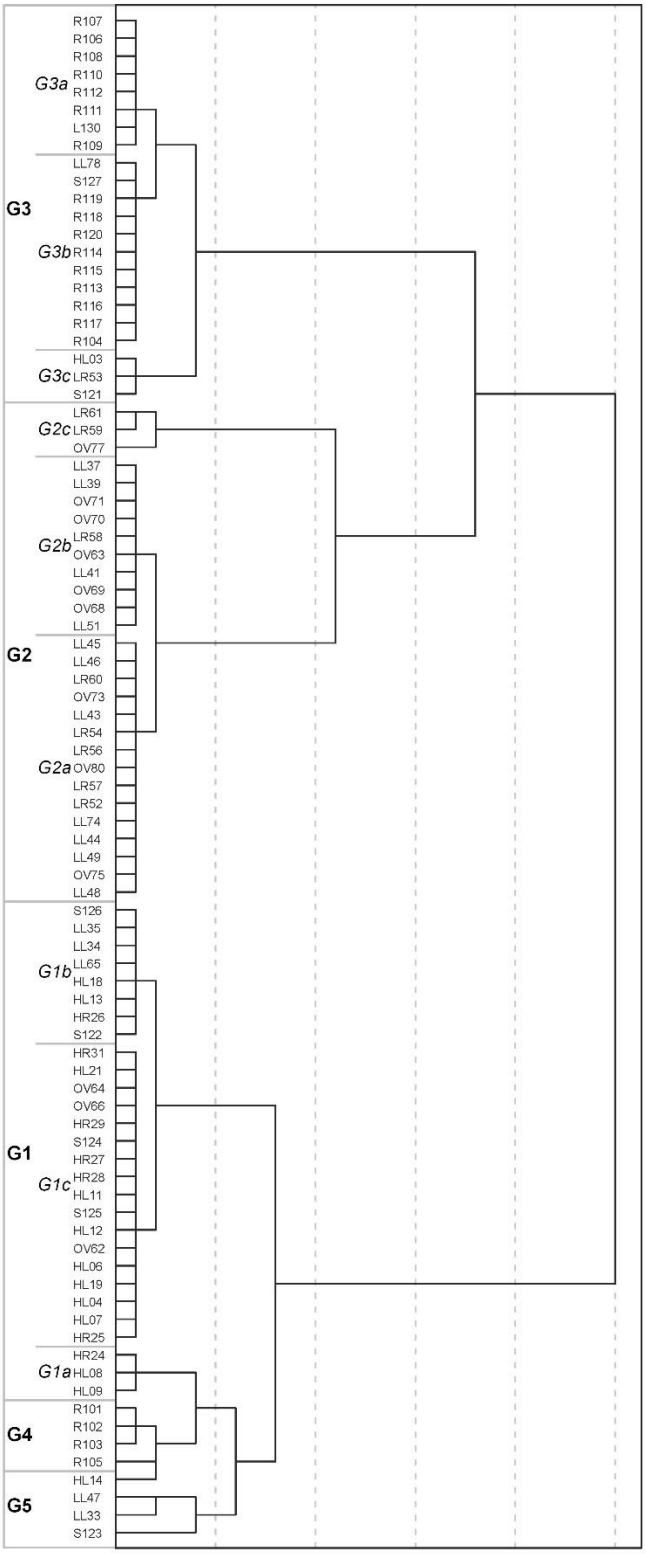



**Figure 2. Dendogram of the cluster analysis showing the different groups and sub-groups identified.**

**Table 1. Number of sampling points and average values of measured parameters for each group and sub-group resulting from the cluster analysis.**

| | G1 | | | | G2 | | | | G3 | | | | G4 | G5 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **TOT** | **G1a** | **G1b** | **G1c** | **TOT** | **G2a** | **G2b** | **G2c** | **TOT** | **G3a** | **G3b** | **G3c** | | |
| *n. samples* | 28 | 3 | 8 | 17 | 28 | 15 | 10 | 3 | 22 | 8 | 11 | 3 | 4 | 4 |
| *pH* | 7.4 | 7.2 | 7.3 | 7.5 | 7.8 | 7.9 | 7.8 | 7.6 | 8.0 | 8.3 | 8.0 | 7.6 | 8.1 | 7.2 |
| *Temp (°C)* | 15.3 | 14.6 | 15.2 | 15.4 | 14.7 | 14.6 | 14.8 | 14.3 | 15.1 | 14.4 | 14.4 | 19.6 | 12.9 | 16.3 |
| *EC (µS/cm)* | 661 | 841 | 741 | 592 | 488 | 487 | 450 | 616 | 364 | 271 | 407 | 458 | 730 | 808 |
| *DO (mg/L)* | 5.46 | 5.46 | 5.43 | 5.47 | 0.34 | 0.10 | 0.81 | 0.02 | 9.02 | 10.56 | 9.18 | 4.37 | 9.82 | 2.00 |
| *Alkalinity (meq/L)* | 5.0 | 6.2 | 5.7 | 4.5 | 4.9 | 5.0 | 4.3 | 6.6 | 2.5 | 1.7 | 2.9 | 3.4 | 5.4 | 6.2 |
| *Cl (mg/L)* | 12.7 | 28.9 | 13.0 | 9.8 | 2.7 | 3.1 | 2.1 | 3.0 | 5.5 | 3.5 | 6.8 | 5.9 | 26.0 | 23.2 |
| *$NO_3$ (mg/L)* | 41.6 | 76.4 | 49.9 | 31.5 | 1.2 | 1.8 | 0.7 | 0.1 | 10.7 | 3.6 | 14.6 | 15.7 | 32.6 | 24.0 |
| *$SO_4$ (mg/L)* | 41.6 | 34.9 | 44.4 | 41.5 | 13.1 | 9.6 | 22.2 | 0.1 | 40.1 | 40.6 | 40.4 | 37.9 | 46.8 | 54.7 |
| *$NH_4$ (mg/L)* | 0.005 | 0.005 | 0.001 | 0.006 | 0.856 | 0.694 | 0.258 | 3.654 | 0.032 | 0.036 | 0.037 | 0.002 | 0.138 | 0.513 |
| *Ca (mg/L)* | 108.0 | 141.5 | 121.2 | 95.9 | 74.6 | 74.0 | 68.5 | 97.7 | 53.9 | 37.8 | 60.2 | 73.7 | 107.7 | 124.4 |
| *Mg (mg/L)* | 14.5 | 18.3 | 16.3 | 13.0 | 12.2 | 12.2 | 12.3 | 11.8 | 8.6 | 7.1 | 9.6 | 8.8 | 16.4 | 15.2 |
| *Na (mg/L)* | 6.0 | 7.5 | 7.3 | 5.1 | 9.7 | 10.4 | 6.9 | 15.4 | 4.5 | 3.1 | 5.2 | 5.4 | 17.7 | 15.2 |
| *K (mg/L)* | 1.4 | 1.3 | 2.0 | 1.2 | 1.4 | 1.3 | 1.5 | 1.7 | 1.4 | 1.2 | 1.6 | 1.4 | 3.7 | 4.4 |
| *As (µg/L)* | 0.6 | 0.5 | 0.7 | 0.6 | 29.8 | 14.5 | 3.1 | 195.2 | 0.9 | 0.7 | 1.1 | 0.6 | 1.8 | 6.0 |
| *Fe (µg/L)* | 4 | <0.1 | 2 | 6 | 482 | 211 | 408 | 2086 | 11 | 18 | 8 | 1 | 24 | 878 |
| *Mn (µg/L)* | 3 | 1 | 2 | 4 | 109 | 90 | 131 | 136 | 6 | 7 | 7 | 2 | 30 | 224 |
| *$\delta^{18}O$ (‰ VSMOW2)* | -8.78 | -7.80 | -8.87 | -8.91 | -8.91 | -8.78 | -9.24 | -8.49 | -9.46 | -9.49 | -9.43 | -9.44 | -8.48 | -8.66 |
| *$\delta^{2}H$ (‰ VSMOW2)* | -58.95 | -50.68 | -59.81 | -60.00 | -59.89 | -58.61 | -62.94 | -56.10 | -64.37 | -65.25 | -63.86 | -63.86 | -56.42 | -58.61 |

The G1 (i.e. higher plain groundwater and springs) is characterized by an oxidized *hydrofacies* with higher EC (average of 661 μS/cm) and $NO_3$ (average of 41.6 mg/L). The higher values of these parameters, that indicate a worse water quality, could be related to anthropogenic activities (mostly agriculture) and to the hydrogeological features of the higher plain aquifer, that is more permeable, and thus, more vulnerable to the existing anthropogenic pressures. The fact that higher plain groundwater and springs fall in the same group agrees with the fact that the spring belt can be considered as a surface discharge of groundwater coming from the higher plain. Within the G1, three sub-groups can be identified: (G1a) groundwaters with the highest $NO_3$ concentrations (average of 76.4 mg/L that exceeds the regulatory limit of 50 mg/L) and a more enriched isotopic signature (average -7.8‰ of $\delta^{18}O$ and -50.68‰ of $\delta^{2}H$) that approaches that of local precipitation (-7.7‰ and -50.10‰ of $\delta^{18}O$ and $\delta^{2}H$, respectively, at Sarnico station; Longinelli and Selmo, 2003); these samples are located in the northern part of the higher plain; (G1b) springs and groundwaters located around the spring belt and characterized by intermediate $NO_3$ concentrations (average of ~50 mg/L) and (G1c) samples characterized by lower $NO_3$ (average of 31.5 mg/L) that are located around the Oglio River.

The G2 (i.e. lower plain groundwater) is characterized by a reduced *hydrofacies* with higher As, Fe, Mn and $NH_4$ that frequently exceed regulatory limits (10, 200, 50 and 500 μg/L, respectively). This hydrochemical feature is related to the hydrogeological properties of the lower plain aquifer system, that has a multi-layer structure, allowing the presence of different overlaying confined aquifers with longer residence time and so older groundwater ages, and contains relevant amounts of buried organic matter as peat sediments. This natural source of organic matter fuels the ecological succession of terminal electron accepting processes allowing the establishment of reducing conditions and high concentrations of their products, as Mn and Fe. The release of As is likely related to the reductive dissolution of Fe-oxides whereas the high concentration of $NH_4$ is directly produced by the organic matter degradation (Rotiroti et al., 2014). Also within the G2, three sub-groups can be identified: (G2a) samples with more reduced states, (G2b) samples with earlier reduced states and (G2c) samples with the highest As concentrations. The G2a contains groundwaters from mostly deep (>100 m b.g.s.) and intermediate (40-100 m b.g.s.) wells, where reducing processes can evolve to more advanced stages (Rotiroti et al., 2015b). Indeed, the G2a has high average concentrations of

As (14.5 μg/L), Mn (90 μg/L), Fe (211 μg/L) and $NH_4$ (694 μg/L) that all exceed the respective regulatory limits. Groundwater samples forming the G2b are located close to the transition between higher and lower plain and/or close to the Oglio River. Therefore, the presence of earlier reduced stages, evidenced by lower As and $NH_4$ and higher Mn, Fe and $SO_4$ with respect to the G2a (see Tab. 1), could be related to shorter groundwater circulation and/or some interactions with surface waters. Since the Oglio River is gaining in this zone, the latter mean recharge from leaking irrigation channels and/or recharge of Oglio River water induced by extensive well pumping, that could locally reverse the natural interrelation between Oglio River and groundwater (i.e., gaining river). The G2c is composed of three groundwater samples with the highest As concentrations (average of 195.2 μg/L). These samples are all located in the south-eastern margin of the study area.

The G3 (i.e. Lake Iseo and Oglio River) is characterized by lower EC (average 364 μS/cm) and a more depleted isotopic signature (average $\delta^{18}O$ and $\delta^{2}H$ of -9.49 and -65.25‰, respectively). The former indicates a general better water quality with respect to groundwater. The latter is due to the fact that Lake Iseo, that feeds the studied stretch of Oglio River, collects waters of Alpine origin that have a more depleted signature (Longinelli and Selmo, 2003). The G3 can be subdivided into 3 sub-groups: (G3a) Lake Iseo and the upstream Oglio River stretch with losing behavior, (G3b) the downstream Oglio River stretch with gaining behavior and (G3c) groundwater and spring directly fed by Oglio River water. The G3a has lower EC (average of 271 μS/cm) and $NO_3$ (average of 3.6 mg/L), confirming the better water quality of Oglio River in its losing stretch. The G3b experiences a decrease of water quality as evidenced by the higher EC (average of 407 μS/cm) and $NO_3$ (average of 14.6 mg/L). This could be the effect of the gaining of groundwater from the higher plain, which has a worse quality, as discussed above. It should be noted that a groundwater sample from the lower plain (LL78) falls within G3b. This could be related to a recharge to the well mainly by surface water, however, a more detailed analysis through new samplings is required to better understand what drives the hydrochemistry of this sample. The G3c consists of two groundwater (HL03 and LR53) and one spring (S121) samples. The well HL03 is located 100 m from the Oglio River in its losing stretch, so a relevant recharge by river water on this well seems reasonable. The well LR53 and the spring S121 are located close to a big irrigation channel fed by Oglio River

water (Fig. 1b). Since most of the irrigation channels leaks water to the aquifer (Facchi et al., 2004), the recharge by river water via irrigation channels on these two monitoring points seems plausible.

The G4 is composed of four of the five tributaries of the Oglio River sampled within the study area. The average values of EC (730 μS/cm) and $NO_3$ (32.6 mg/L) in G4 are higher with respect to those of G3 (364 μS/cm and 10.7 mg/L, respectively) indicating that tributary rivers have a general worse water quality with respect to Oglio River itself. The only one tributary sample that is out from G4 (R104) is from a channel named Scolmatore di Genivolta, that is fed by irrigation water sourced by Oglio and Adda rivers. The fact that this channel is fed by Oglio River, together with the absence of anthropogenic pressures on it, leads the sample R104 to be grouped into G3b rather than G4.

The G5 collects four groundwater and spring samples (HL14, LL33, LL47 and S123) that can be considered as outliers. The samples HL14 and LL33 have respectively a higher Cl concentration (43.9 mg/L) and EC value (879 μS/cm) that could be related to local anthropogenic influences. The sample LL47 has high As, Fe and Mn concentrations (higher than regulatory limits) and, at the same time, higher EC, Cl and $SO_4$ (with respect to the other lower plain groundwater samples). This could be due to a mixing in the well of groundwater coming from different overlying aquifers with different hydrochemical features. The spring S123 has a more reduced *hydrofacies* (DO of 0.04 mg/L and Mn of 194 μg/L) with respect to the other spring samples and the highest measured values of K (10.8 mg/L). This could be due to a particular recharge system of this spring that could likely involve reduced groundwater from lower plain aquifers together with possible anthropogenic influences.

Results of FA generally confirm the sample grouping derived from the CA. More specifically, results of FA showed that only the first four factors have an eigenvalue higher than 1, and thus, can be considered as significant factors. These explain the 79.7% of the variance of the original dataset (cumulative explained variance). The F1 and F2 explain the 37.9% and 23.2% of the variance, respectively, whereas the F3 and F4 explain lower percentages of variance, that are 11.3% and 7.4%, respectively.

The Fig. 3 shows the loading and the score plots of F1 *vs* F2. Concerning the loadings of F1, the most important original variables result EC, alkalinity, Mg, Ca, water isotopes, Cl and pH, the latter is negatively correlated to the other parameters. In general, this factor seems to represent major ions and water isotopes. Concerning the loadings of F2, its most important original variables are associated into two groups that are negatively correlated each other: a) the first group is formed by $NH_4$, As, Fe and Mn, b) the second group involves DO, $NO_3$ and $SO_4$. Therefore, the F2 likely represents the redox conditions of water samples.

In the score plot of Fig. 3b, the sampling points are classified for each group resulted from the CA. In general, each group and sub-group identified in the CA is also well identifiable in the score plot of F1 *vs* F2 confirming the strength of the hydrochemical characterization performed. Along the F1, a gradual evolution from surface waters (G3) with higher pH and lower EC and major ions to higher plain groundwater (G1) with higher EC and $NO_3$ is pointed out. Moreover, along the F2, the evolution from the oxidized *hydrofacies* of surface waters (G3) and higher plain groundwater (G1) to the reduced *hydrofacies* of lower plain groundwater (G2) is also well represented.

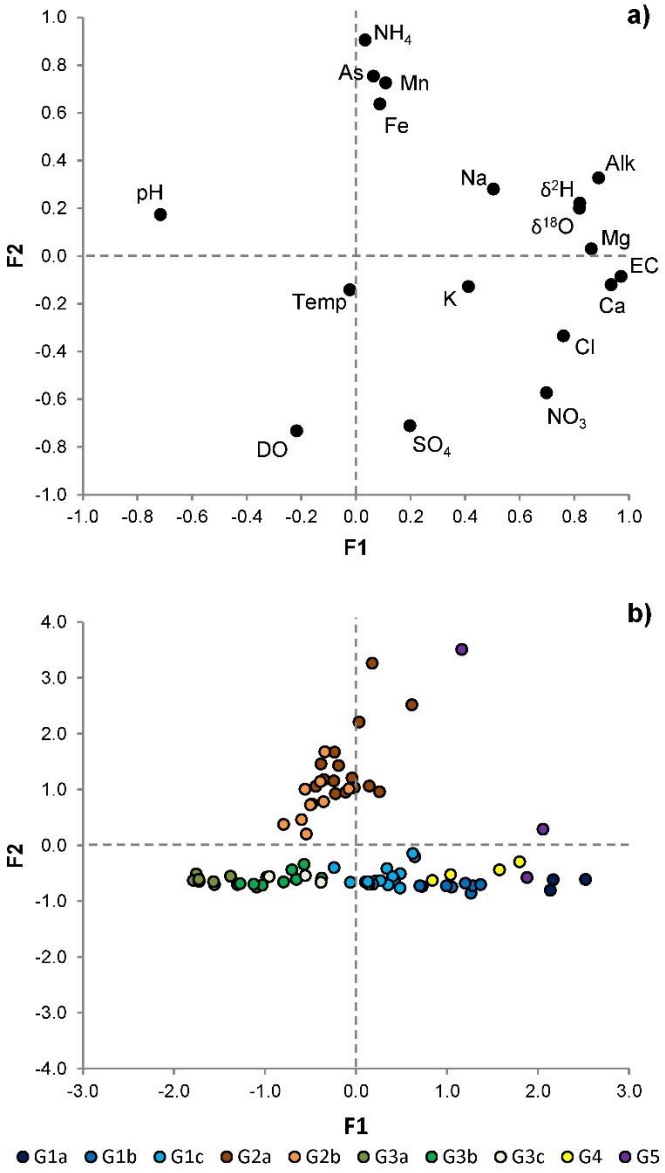



**Figure 3. a) Loading plot of F1 *vs* F2. b) Score plot of F1 *vs* F2, colors represent the different groups and sub-groups resulting from the cluster analysis.**

### 2.2.3 Conclusions

This work presented a joined hydrochemical characterization of either surface water and groundwater based on multivariate statistical analysis, such as cluster analysis and factor analysis.

Results pointed out that Lake Iseo and Oglio River, higher plain groundwater and springs, lower plain groundwater and Oglio River tributaries form 4 discrete groups with particular distinctive features. Each group can be subdivided into up to 3 sub-groups on the basis of different hydrodynamic, hydrogeological and hydrochemical features.

In general, results confirm how multivariate statistical analysis can support the interpretation of large hydrochemical datasets. The main advantage of using this technique for interpreting hydrochemical data with respect to "traditional" methods, such as Piper diagram and bivariate plots, is that multivariate statistics are independent and quantitative methods that are able to extract information from all available data by elaborating jointly and simultaneously all measured variables and samples. In the present work, results of multivariate statistical analysis led to

improve the hydrological and hydrogeological conceptual model of the area and elucidate the effects on water chemistry of groundwater/surface water interactions. The improvement of the conceptual model will support the future implementation of a numerical flow modeling of the studied system.

### 2.2.4 Acknowledgements

This work was supported by Fondazione Cariplo, grant 2014-1282.

### 2.2.5 References

Angelone M., Cremisini C., Piscopo V., Proposito M. & Spaziani F. (2009) - Influence of hydrostratigraphy and structural setting on the arsenic occurrence in groundwater of the Cimino-Vico volcanic area (central Italy). Hydrogeol. J., 17, 901-914. DOI:10.1007/s10040-008-0401-3

Bonfanti M., Ducci D., Masetti M., Sellerino M. & Stevenazzi S. (2016) - Using statistical analyses for improving rating methods for groundwater vulnerability in contamination maps. Environ. Earth Sci., 75, 1003. DOI:10.1007/s12665-016-5793-0

Bonomi T. (2009) - Database development and 3D modeling of textural variations in heterogeneous, unconsolidated aquifer media: Application to the Milan plain. Comput. Geosci., 35, 134-145. DOI:10.1016/j.cageo.2007.09.006

Buizza M. (eds.) (2012) - Progetto di sperimentazione dei rilasci per il deflusso minimo vitale nel fiume Oglio sub lacuale - Relazione conclusiva del triennio. Consorzio dell'Oglio, Brescia, 102 pp.

Capri E., Civita M., Corniello A., Cusimano G., De Maio M., Ducci D., Fait G., Fiorucci A., Hauser S., Pisciotta A., Pranzini G., Trevisan M., Delgado Huertas A., Ferrari F., Frullini R., Nisi B., Offi M., Vaselli O. & Vassallo M. (2009) - Assessment of nitrate contamination risk: The Italian experience. J. Geochem. Explor., 102, 71-86. DOI:10.1016/j.gexplo.2009.02.006

Carraro A., Fabbri P., Giaretta A., Peruzzo L., Tateo F. & Tellini F. (2015) - Effects of redox conditions on the control of arsenic mobility in shallow alluvial aquifers on the Venetian Plain (Italy). Sci. Total Environ., 532, 581-594. DOI:10.1016/j.scitotenv.2015.06.003

Dalla Libera N., Fabbri P., Mason L., Piccinini L. & Pola M. (2017) - Geostatistics as a tool to improve the natural background level definition: An application in groundwater. Sci. Total Environ., 598, 330-340. DOI:10.1016/j.scitotenv.2017.04.018

Delconte C.A., Sacchi E., Racchetti E., Bartoli M., Mas-Pla J. & Re V. (2014) - Nitrogen inputs to a river course in a heavily impacted watershed: A combined hydrochemical and isotopic evaluation (Oglio River Basin, N Italy). Sci. Total Environ., 466–467, 924-938. DOI:10.1016/j.scitotenv.2013.07.092

Éupolis Lombardia. (2015) - Progetto di accompagnamento a supporto del processo di revisione del piano di tutela delle acque - attività di progettazione, monitoraggio e studio relative ai corpi idrici sotterranei della lombadia - Relazione di Sintesi, 24 pp.

Facchi A., Ortuani B., Maggi D. & Gandolfi C. (2004) - Coupled SVAT–groundwater model for water resources simulation in irrigated alluvial plains. Environ. Model. Software, 19: 1053-1063. DOI:10.1016/j.envsoft.2003.11.008

Facchinelli A., Sacchi E. & Mallen L. (2001) - Multivariate statistical and GIS-based approach to identify heavy metal sources in soils. Environ. Pollut., 114, 313-324. DOI:10.1016/S0269-7491(00)00243-8

Gibrilla A., Bam E.K.P., Adomako D., Ganyaglo S., Osae S., Akiti TT, Kebede S., Achoribo E., Ahialey E., Ayanu G. & Agyeman E. K. (2011) - Application of Water Quality Index (WQI) and Multivariate Analysis for Groundwater Quality Assessment of the

Birimian and Cape Coast Granitoid Complex: Densu River Basin of Ghana. Water Qual. Expo. Health, 3, 63. DOI:10.1007/s12403-011-0044-9

Kaiser H.F. (1958) - The varimax criterion for analytic rotation in factor analysis. Psychometrika, 23, 187-200. DOI:10.1007/bf02289233

Lasagna M., De Luca D.A. & Franchino E. (2016) - Nitrate contamination of groundwater in the western Po Plain (Italy): the effects of groundwater and surface water interactions. Environ. Earth Sci., 75, 240. DOI:10.1007/s12665-015-5039-6

Leoni B., Garibaldi L. & Gulati R.D. (2014) - How does interannual trophic variability caused by vertical water mixing affect reproduction and population density of the Daphnia longispina group in Lake Iseo, a deep stratified lake in Italy? Inland Waters, 4, 193-203. DOI:10.5268/iw-4.2.663

Longinelli A. & Selmo E. (2003) - Isotopic composition of precipitation in Italy: a first overall map. J. Hydrol., 270, 75-88. DOI:10.1016/S0022-1694(02)00281-0

Mantovi P., Fumagalli L., Beretta G.P. & Guermandi M. (2006) - Nitrate leaching through the unsaturated zone following pig slurry applications. J. Hydrol., 316, 195-212. DOI:10.1016/j.jhydrol.2005.04.026

Mastrocicco M., Colombani N. & Gargini A. (2014) - Modelling present and future Po river interactions with alluvial aquifers (Low Po River Plain, Italy). J. Water Clim. Change, 5, 457-471. DOI:10.2166/wcc.2014.058

Palmucci W., Rusi S. & Di Curzio D. (2016) - Mobilisation processes responsible for iron and manganese contamination of groundwater in Central Adriatic Italy. Environ. Sci. Pollut. Res., 23, 11790-11805. DOI:10.1007/s11356-016-6371-4

Perego R., Bonomi T., Fumagalli L., Benastini V., Aghib F., Rotiroti M. & Cavallin A. (2014) - 3D reconstruction of the multi-layer aquifer in a Po Plain area. Rend. Online Soc. Geol. It., 30, 41-44. DOI:10.3301/ROL.2014.09

Re V., Sacchi E., Mas-Pla J., Menció A. & El Amrani N. (2014) - Identifying the effects of human pressure on groundwater quality to support water management strategies in coastal regions: A multi-tracer and statistical approach (Bou-Areg region, Morocco). Sci. Total Environ., 500, 211-223. DOI:10.1016/j.scitotenv.2014.08.115

Rotiroti M., Bonomi T., Fumagalli L., Taviani S., Stefania G.A., Zanotti C., Patelli M., Soler V. & Leoni B. (2016) - Interactions between river water and groundwater and their

influences on river chemistry: the case study of Oglio River (northern Italy). Proceedings of 1° Congresso Nazionale Congiunto SITE - UZI - SIB, Milano, 30 agosto - 2 settembre 2016.

Rotiroti M., Di Mauro B., Fumagalli L. & Bonomi T. (2015a) - COMPSEC, a new tool to derive natural background levels by the component separation approach: application in two different hydrogeological contexts in northern Italy. J. Geochem. Explor., 158, 44-54. DOI:10.1016/j.gexplo.2015.06.017

Rotiroti M., Jakobsen R., Fumagalli L. & Bonomi T. (2015b) - Arsenic release and attenuation in a multilayer aquifer in the Po Plain (northern Italy): Reactive transport modeling. Appl. Geochem., 63, 599-609. DOI:10.1016/j.apgeochem.2015.07.001

Rotiroti M., McArthur J., Fumagalli L., Stefania G.A., Sacchi E. & Bonomi T. (2017) - Pollutant sources in an arsenic-affected multilayer aquifer in the Po Plain of Italy: Implications for drinking-water supply. Sci. Total Environ., 578, 502-512. DOI:10.1016/j.scitotenv.2016.10.215

Rotiroti M., Sacchi E., Fumagalli L. & Bonomi T. (2014) - Origin of arsenic in groundwater from the multi-layer aquifer in Cremona (northern Italy). Environ. Sci. Technol., 48: 5395–5403. DOI:10.1021/es405805v

Sacchi E., Acutis M., Bartoli M., Brenna S., Delconte C.A., Laini A. & Pennisi M. (2013) - Origin and fate of nitrates in groundwater from the central Po plain: Insights from isotopic investigations. Appl. Geochem., 34, 164-180. DOI:10.1016/j.apgeochem.2013.03.008

Valerio G., Pilotti M., Barontini S. & Leoni B. (2015) - Sensitivity of the multiannual thermal dynamics of a deep pre-alpine lake to climatic change. Hydrol. Process., 29: 767-779. DOI:10.1002/hyp.10183

Ward J.H. (1963) - Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc., 58, 236-244. DOI:10.1080/01621459.1963.10500845

## 2.3 Groundwater and surface water quality characterization through positive matrix factorization combined with GIS approach.

Zanotti, C.[1], Rotiroti, M. [1], Fumagalli, L. [1], Stefania, G.A. [1], Canonaco, F. [2], Stefenelli, G. [2], Prévôt, A.S.H. [2], Leoni, B [1], Bonomi, T. [1].

[1] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza, 1, 20126 Milano, Italy. Email: c.zanotti@campus.unimib.it

[2] Laboratory of Atmospheric Chemistry, Paul Scherrer Institute, 5232 Villigen-PSI, Switzerland

**Water Research Vol. 159 (2019), pp 122 – 134**

**Doi:** https://doi.org/10.1016/j.watres.2019.04.058

### 2.3.1 Abstract

This study aims at testing the effectiveness of Positive Matrix Factorization in characterizing groundwater and surface water quality, in terms of identifying main hydrochemical features and processes (natural and anthropogenic) that govern them. This method is applied in a hydro-system featured by a strong interrelation between groundwater and surface water and highly impacted by agricultural activities. Therefore, a holistic approach considering groundwater together with the surface water bodies, consisting in lake, several rivers and springs, was used.

Multivariate statistical analysis, in particular Factor Analysis, has been proved to be effective in elaborating and interpreting water quality data highlighting the information carried within them, but it presents some limitations: it does not consider data uncertainty and it groups variables which are correlated positively and negatively. Moreover, in some cases the resulting factors are not clearly interpretable, describing each one various overlapping features/processes.

Here, Positive Matrix Factorization is applied to groundwater and surface water quality data, and the results are compared to those obtained through a Factor Analysis in terms of both factor profiles and their spatial distribution through a GIS approach. Results of isotopes analysis are used to validate PMF output and support interpretation. Positive Matrix

Factorization allows to consider data uncertainty and the solution respects two positivity constraints, based on the concept of chemical mass balance, which leads to a more environmentally interpretable solution.

Results show that Positive Matrix Factorization identifies five different factors reflecting main features and natural and anthropogenic processes affecting the study area: 1) surface water used for irrigation, 2) groundwater subjected to reducing processes at advanced stages, 3) groundwater subjected to reducing processes at early stages, 4) groundwater residence time and 5) the effects of the agricultural land use on both groundwater and surface water.

Positive Matrix Factorization leads to a more detailed understanding of the studied system as compared to Factor Analysis which identifies only three factors with overlapping information. Based on the results of this study, Positive Matrix Factorization could be a useful technique to perform groundwater and surface water quality characterization and to reach a deeper understanding of the phenomena that govern water chemistry.

KEYWORDS: multivariate statistical analysis, water quality, Positive Matrix Factorization, Factor Analysis, Oglio River

### 2.3.2 Introduction

Groundwater and surface water are the main sources of freshwater and drinking water in many regions worldwide. Their quality should be preserved in order to guarantee the fulfilment of the different water needs by humans and ecosystems. Groundwater and surface water quality could be affected by several factors and processes, such as anthropogenic impacts and natural phenomena as well as groundwater/surface-water interrelationships. The understanding of the different water and pollution sources and the processes that affect them is fundamental to obtain an exhaustive characterization of water resources quality. To this end, the implementation of a proper water quality monitoring network is fundamental, but it is also crucial to properly manage, elaborate and interpret the collected data in order to gather all the information contained within them.

In the last decades, attention toward multivariate statistical analysis increased dramatically as a tool to analyse water chemical data. Several authors performed Principal Component analysis (PCA), Factor Analysis (FA) or Cluster Analysis (CA) to investigate groundwater and surface water chemical data and identify natural phenomena (Blake et al., 2016; Koh et al., 2016; Shrestha et al., 2016) as well as anthropogenic impacts affecting water quality (Alberti et al., 2016; Devic et al., 2014; Gu et al., 2018; Phung et al., 2015; Stefania et al., 2018).

Positive Matrix Factorization (PMF) (Paatero and Tapper, 1994) is a multivariate analysis aimed at source identification and apportionment, specifically designed to cope with environmental data and manage their uncertainty and distributions. PMF is specifically appropriate for environmental data because: (1) it takes into account the analytical uncertainties often associated with measurements of environmental samples and (2) forces all of the values in the solution profiles and contributions to be positive, which can lead to a more realistic representation of the system than solutions from other multivariate methods like PCA (Reff et al., 2007).

In the last two decades, PMF has been widely used in studies concerning air pollution, since it was specifically designed to cope with application where mass balance is a key aspect and a known source is present. Particularly, PMF has been widely applied to perform source apportionment of particulate matter and aerosol in several countries (e.g. Belis et al., 2011; Bozzetti et al., 2017; Daellenbach et al., 2017; Fortner et al., 2018; Kuang et al., 2015; Mohr et al., 2012; Sowlat et al., 2016; Visser et al., 2015; Yan et al., 2016).

Only in the last years, a few authors applied PMF on datasets concerning different environmental compartments to reach a more realistic representation of the sources affecting different systems. Comero et al. (2011) applied PMF on a geochemical dataset on sediments from 11 alpine lakes located in Italy, identifying four interpretable factors related to the mineralogical/chemical features of lake sediments in the catchment area. Comero et al. (2012) used PMF on soil sample on an abandoned mine dump in Italy and supported the interpretation of the sources with a GIS approach. Zheng et al. (2014) performed a comprehensive study investigating with PMF the presence of Polycyclic Aromatic Hydrocarbons (PAHs) in soils samples. Shao et al. (2014) used PMF model to better understand the PAHs sources in a karst water system. Soonthornnonda et al. (2008) used

chemical mass balance and PMF modelling in order to find origins of flows and pollutants in combined sewer wastewater. Pekey et al. (2013) investigated through PMF analysis sources affecting surface sediment samples in a bay in Turkey. Gholizadeh et al. (2016) identified and quantified the potential pollution sources affecting the water quality of three major rivers of South Florida by means of PMF.

The low uptake of PMF in water research compared to air pollution field is due to the lower suitability of water quality data to fit the requirements of the technique. Some limitations that could limit a successful PMF analysis are, for instance: 1) PMF requires only data expressed as concentration while some typical measurement of water samples have different units (e.g. pH, Electrical Conductivity, Oxidation Reduction Potential, isotopes analysis, age tracers) thus they cannot be directly fed into a PMF model, 2) a proper monitoring network, specifically designed to capture the variability of the system and its contamination sources, is crucial to obtain a complete representation of the system with a PMF analysis and 3) in cases where a single source is present a multivariate statistical analysis such as PMF might not be appropriate. Despite these limitations, the potential advantages of PMF (i.e. the inclusion of analytical uncertainties and the positivity constraints) could be useful when investigating groundwater and surface water quality dataset to perform water quality characterization.

The aim of this work is to test the effectiveness of PMF as a tool to support water quality characterization intended to define the main hydrochemical features, pollution phenomena and processes governing them, such as groundwater/surface water interactions. Here PMF is applied to a dataset concerning groundwater, spring, river and lake water collected within the Oglio River basin, located in the Po Plain (N Italy). This area was subjected to previous studies (Rotiroti et al. 2019a, 2019b) which identified main hydrochemical features and processes governing them, so the information obtained through the PMF analysis can be tested with a known conceptual model of the study area. The results of the PMF are then compared with those of a "traditional" multivariate statistical analysis such a FA, in order to make a comparison with a more widely used multivariate technique.

To the best of the authors' knowledge, PMF has never been applied to characterize the water quality of a whole hydro-system using a dataset consisting in groundwater and surface water samples.

Moreover, additional innovative aspects of this work are: 1) both factor contributions to the samples and the residuals of the model are spatially analysed through GIS approach and their interpretation is supported by considering the land use and hydrogeological features; 2) results of the PMF model are validated through a comparison with water isotope data.

## 2.3.3 Materials and methods

### *2.3.3.1 Study area*

The study area is the part of the Oglio River basin (Po Plain, N Italy) that covers ~2000 km2 around the Oglio River between the outflow from Lake Iseo and the confluence with Mella River (Figure 1). In this part of its course the Oglio River receives water from 5 tributaries which are: the Cherio River, the Scolmatore di Genivolta channel, the Saverona Stream, the Strone River and the Mella River (Figure 1).

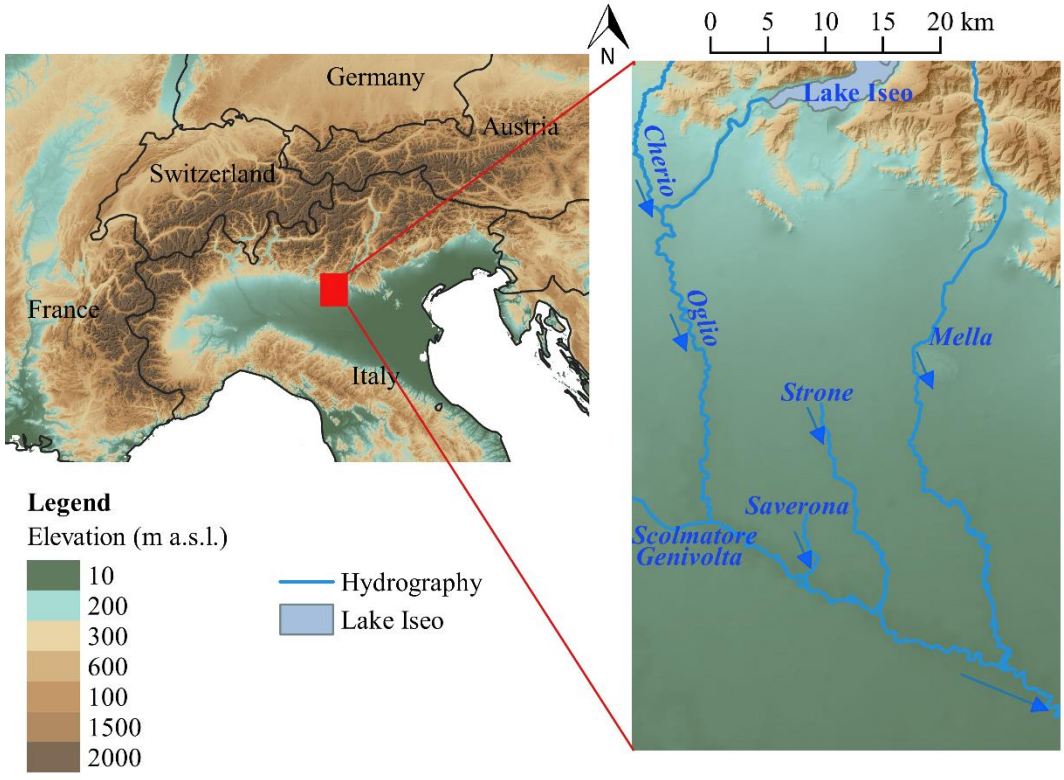



**Figure 1 - location of the study area.**

As regards the geomorphology (Figure 2a) and hydrogeology, the study area can be divided into two parts with different characteristics (Rotiroti et al. 2019a, 2019b): 1) the higher plain in the northern part that hosts a mono-layer unconfined aquifer mainly composed of sands and gravels and 2) the lower plain in the southern part of the study area that hosts a multi-layer aquifer where several layers of sands and silty clays alternate vertically. This hydrogeological setting features the entire Alpine sector of the Po Plain (Giuliano, 1995; Perego et al., 2014). The transition zone between the higher and the lower plain is known as "the springs belt", as natural groundwater outflows are largely present all over the width of the area.

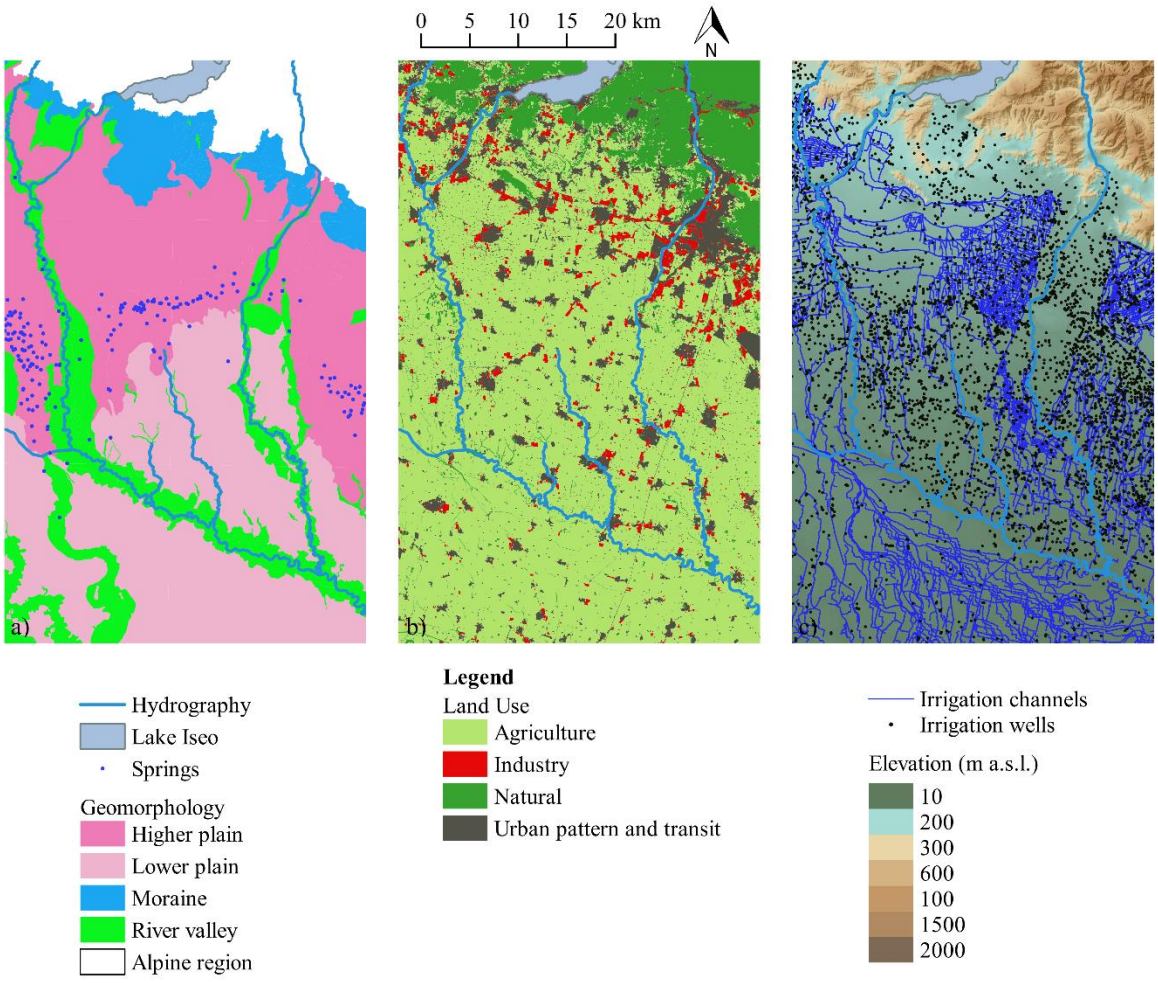



**Figure 2 - a) geological map of the study area; b) land use map; c) irrigation channels and wells.**

Groundwater flows mainly from north to south (Rotiroti et al. 2019a). As regards the relationship between Oglio River and groundwater, the first stretch of Oglio River (i.e. from the outflow from Lake Iseo to approximately 20-30 km downstream) is losing, then it becomes gaining up to the confluence with Mella River (Rotiroti et al. 2019a).

Lake Iseo (Site LTER_EU_IT_008 - "Southern Alpine Lakes"; www.lter-europe.net) is the fourth largest Italian lake. It has been classified as 'warm monomictic', water temperatures do not drop below 4°C and water mixing occurring during or close to winter. However, due to their morphological characteristics and climatic conditions, the lake does not reach a full thermal and chemical homogenisation every year and the complete water overturn over the last 40 years has happened only occasionally (in 2005 and 2006) (Leoni et al., 2014a; Pareeth et al., 2017), so the lake can presently be regarded as holo-oligomictic. In the last five years, the depth of mixing range between 30 and 75 m below the surface, strongly influencing the physical and chemical features of epilimnion and of outlet waters. Lake Iseo has been affected by an increase in mean total phosphorus concentrations since the 1970s. Phosphorus concentrations gradually increased in water column after 2005, from about 60 to 90–100 μg/L, and in epilimnion lightly decrease from 38 to 18 μg/L during the last five years (Leoni et al., 2014b; Marti et al., 2016; Rogora et al., 2018).

The land use of the area is mainly agricultural (Figure 2b). The most frequent crop is corn, mainly cultivated for the purpose of animal husbandry feeding (cattle and pig). Within the area two different irrigation techniques are used respectively in the higher and the lower plain. In the former, irrigation water mainly comes from the Oglio River through a complex network of irrigation channels, in the latter the main source of water for irrigation is groundwater, abstracted through hundreds of irrigation wells (Figure 2c). In the higher plain, that is formed by coarse sediments and thus has higher permeability, irrigation water sourced by the Oglio River constitutes a main source of aquifer recharge (Rotiroti et al. 2019a).

As regards groundwater quality, two main problems affect the study area (Rotiroti et al. 2019a): a) a diffuse nitrate contamination in the higher plain, such as in most of the higher part of the whole Po Plain (Cinnirella et al., 2005; Lasagna et al., 2016; Martinelli et al., 2018) and b) the natural contamination by Mn, Fe, As and NH4 in the lower plain, such as in most of the lower part of the whole Po Plain (Carraro et al., 2015; Molinari et al., 2012; Rotiroti et al., 2014). The

former is mainly related to the agricultural land use and particularly to fertilization practices, although point sources from sewage systems can be relevant in some cases (Sacchi et al 2013; Rotiroti et al. 2017). The latter is related to the hydrogeological and hydrochemical properties of the lower plain aquifer system that is multi-layered and contains relevant amounts of buried organic matter as peat sediments. These feed the ecological succession of terminal electron accepting processes leading to reducing conditions and high concentrations of Mn, Fe, As and NH4. (Rotiroti el al 2015; Carraro et al 2015).

Concerning the water quality of surface water, Lake Iseo has a better water quality as compared to the Oglio River which has higher NO3 concentrations due to the gaining of groundwater from the higher plain that is affected by NO3. Spring water is also affected by NO3 since the spring belt constitutes, in fact, a surface outflow of groundwater from the higher plain. Tributaries of Oglio River have generally worse water quality with respect to the Oglio River itself, being more impacted by anthropogenic activities (Rotiroti et al. 2019a).

#### *2.3.3.2 Available data*

The available dataset is the result of four field surveys conducted in the study area during February 2016, June 2016, September 2016 and March 2017. In each campaign samples from 68 monitoring points were collected including a total amount of 270 samples. Among the monitoring points 44 are wells, 17 are rivers (12 stations on the Oglio River and 5 tributaries), 6 are springs and 1 is the Lake Iseo.

Each sample was analysed for dissolved oxygen (DO), total phosphorous (P-tot), alkalinity as HCO3, major ions (Ca, Mg, Na, K, Cl, SO4, NO3 and NH4), trace elements (As, Fe and Mn) and stable water isotopes (δ18O and δ2H).

As regards water sampling from wells, they were purged until physico-chemical parameters were stable, generally after 2−3 well volumes were removed. The DO was measured in the field using the WTW Multi 3430 meter in a flow cell.

Each water sample was filtered through 0.2 μm filters on the field and samples for As, Fe and Mn analysis were acidified with nitric acid. After collection, samples were stored in a portable fridge at 4°C. Alkalinity was analysed by H2SO4 titration within 24 hours from the sampling.

Major ions were analysed by ion chromatography. Ammonium was analysed by spectrophotometry with Nessler's reagent within 24 hours from the sampling. Iron and manganese were analysed by Inductively Coupled Plasma - Optical Emission Spectroscopy (ICP-OES) whereas arsenic was analysed by Graphite Furnace Atomic Absorption Spectrometry (GFAAS). Water isotopes were analysed by wavelength-scanned cavity ring-down spectroscopy (WS-CRDS). The analytical uncertainty of the data was calculated for each set of analysed samples. Particularly, before every set of analysis a standard solution was measured. The analytical uncertainty was calculated as the percent difference between the value of a standard solution and the measurement with the analytical instruments.

Furthermore, since previous studies highlighted that field sampling operations can give a significant contribution to the total uncertainty (Grøn et al., 2007; Witczak, 2006) with values up to 18% (Roy and Fouillac, 2004), the sampling uncertainty was also accounted. The uncertainty related to the sampling, transport and stocking of the samples was assumed to be 10%. This value was chosen according with the average variation of measured concentrations (14%) between two consecutive years in the same season (i.e. winter 2016 and winter 2017), considered here as a proxy of the uncertainty related to sampling operations. Analytical and field sampling uncertainty were combined according with the recommendations of the EU project BRIDGE (Witczak, 2006). The main statistics of the data and their combined uncertainty are reported in Table 1.

The resulting concentration dataset, used as input for the PMF together with the uncertainty dataset, is a data matrix consisting in 270 samples described by 14 variables (DO, P-tot, alkalinity as HCO3, Ca, Mg, Na, K, Cl, SO4, NO3, NH4, As, Fe and Mn). Water isotopes were excluded from PMF since they are not conventionally expressed as a mass concentration and thus they cannot be directly inserted as an input of the PMF analysis. However, water isotopes were used for a validation of the PMF results.

**Table 1 – main statistics of data and uncertainty used as PMF input**

| | data | | | | | uncertainties | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | min | max | average | median | std dev | min | max | average | median | std dev |
| $O_2$ (mg/L) | <0.001 | 14.04 | 4.44 | 4.37 | 4.27 | - | 1.40 | 0.45 | 0.44 | 0.43 |
| P-tot (µg/L) | 0.87 | 1042.20 | 118.16 | 55.09 | 170.86 | 0.09 | 161.35 | 14.76 | 5.79 | 23.63 |
| Alcalinity (mg/L) as $HCO_3$ | 84.59 | 435.15 | 273.46 | 286.01 | 75.06 | 8.46 | 43.52 | 27.35 | 28.60 | 7.51 |
| Cl (mg/L) | 0.95 | 40.21 | 8.30 | 6.35 | 6.99 | 0.10 | 4.08 | 0.84 | 0.65 | 0.71 |
| $NO_3$ (mg/L) | <0.1 | 99.04 | 17.20 | 10.52 | 19.80 | - | 10.07 | 1.75 | 1.09 | 2.01 |
| $SO_4$ (mg/L) | <0.1 | 61.35 | 29.46 | 37.72 | 17.50 | - | 6.23 | 3.00 | 3.83 | 1.79 |
| $NH_4$ (mg/L) | <0.001 | 5.06 | 0.47 | 0.12 | 0.87 | - | 0.97 | 0.06 | 0.01 | 0.12 |
| Ca (mg/L) | 34.41 | 144.16 | 83.23 | 78.27 | 24.12 | 3.46 | 14.47 | 8.35 | 7.85 | 2.42 |
| Mg (mg/L) | 6.61 | 21.89 | 12.51 | 12.40 | 3.02 | 0.66 | 2.21 | 1.26 | 1.25 | 0.31 |
| Na (mg/L) | 2.63 | 20.79 | 7.98 | 7.33 | 3.45 | 0.26 | 2.09 | 0.80 | 0.74 | 0.35 |
| K (mg/L) | 0.74 | 10.37 | 1.75 | 1.43 | 1.14 | 0.08 | 1.07 | 0.18 | 0.15 | 0.12 |
| As (µg/L) | <0.05 | 289.31 | 14.36 | 0.89 | 42.38 | - | 39.17 | 1.75 | 0.12 | 5.27 |
| Fe (µg/L) | <0.1 | 3270.00 | 209.55 | 13.50 | 480.34 | - | 486.12 | 28.14 | 1.62 | 66.32 |
| Mn (µg/L) | <0.1 | 267.00 | 47.68 | 11.16 | 59.54 | - | 54.98 | 7.71 | 1.85 | 10.36 |

#### *2.3.3.3 Positive Matrix Factorization*

PMF (Paatero and Tapper, 1994) is a multivariate analysis in which the fundamental problem is to resolve the identities and contributions of different sources in a mixture. The bilinear factorises a multivariate dataset into two matrices, G and F, leading to a reproduction of the dataset variability as a linear combination of a set of constant factors profile and their contribution to each sample.

Unlike FA or PCA, and similarly to other techniques (e.g. Parallel Factor Analysis, PARAFAC), in the PMF model no constraint are imposed about the orthogonality of the factors. On the other hand, other constraints are considered, aimed at reproducing natural physical limit of the system which are: a) the composition of predicted source must be positive (a source cannot have a negative contribution of an element); b) the predicted source contribution to each sample must all be positive (a source cannot contribute to a sample with a negative mass).

The solution of the problem is found through a weighted least squares approach, aimed at minimizing the value of the object function $Q$ for a given number of factor $p$. $Q$ is defined as:

$$Q = \sum_{i=1}^{n} \sum_{i=1}^{m} \left( \frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{\sigma_{ij}} \right)^2 \quad (1)$$

Where $\sigma_{ij}$ is the uncertainty of species j concentration data in the $i_{th}$ sample, $m$ is the number of analytes and $n$ is the number of samples.

It transpires that the solution to the PMF problem depends on estimating uncertainties for each of the data points of the PMF input (Hopke, 2000). In environmental datasets three types of data are typically available: 1) in most of cases samples have concentration value that has been determined, $x_{ij}$, and their associated uncertainty $\sigma_{ij}$; 2) there are samples in which certain species could not be precisely determined because their concentration was below the detection limit; 3) "missing data" which have not been measured. In the present study the following method to determine concentrations value and their associated error estimates was used according to Polissar et al. (1998) and Reff et al. (2007):

$$x_{ij}^k = c_{ij}^k \qquad \text{for determined values}$$

$$x_{ij}^k = {d_{ij}^k}/{2} \qquad \text{for values below detection limit} \quad (2)$$

$$x_{ij}^k = \tilde{c}_{ij}^k \qquad \text{for missing values}$$

$$\sigma_{ij}^k = u_{ij}^k + {d_{ij}^k}/{3} \qquad \text{for determined values}$$

$$\sigma_{ij}^k = \frac{5}{6} d_{ij}^k \qquad \text{for values below detection limit} \quad (3)$$

$$\sigma_{ij}^k = 4\tilde{c}_{ij}^k \qquad \text{for missing values}$$

where $c_{ij}^k$, $u_{ij}^k$ and $d_{ij}^k$ are, respectively, the measured concentration, the analytical uncertainty, and the method detection limit for sample *i,* element *j*, and sampling site *k*. $\tilde{c}_{ij}^k$ is the geometric mean of the measured concentration of the element *j* at sampling site *k*.

To explore the reliability of different variables and different samples the signal-to-noise (S/N) value was considered, which is the ratio between the concentration and its uncertainty. Data with a low signal to noise (i.e. S/N between 0.2 and 2) are considered "weak", while data with a S/N lower than 0.2 are considered "bad". In this work weak and bad concentrations values have been downscaled during the PMF analysis according to Paatero and Hopke 2003.

Concerning to the data weighting, another important aspect to consider is the weight of extreme values. Environmental data typically show a skewed distribution with a heavy tail due to the presence of samples with higher concentration of certain species as compared to other samples. These data can have a significant leverage on the solution of the model (Comero et al., 2009), thus it is possible run the model in the so called "robust mode" where the extreme values influence is reduced. In the robust mode the objective function is changed as follows:

$$Q_{Robust} = \sum_{i=1}^{n} \sum_{i=1}^{m} \left( \frac{e_{ij}}{h_{ij}\sigma_{ij}} \right)^2 \quad (4)$$

Where

$$h_{ij} = 1 \quad \text{for} \quad \left| e_{ij}/\sigma_{ij} \right| \leq \alpha$$

$$h_{ij} = \left| e_{ij}/\sigma_{ij} \right| / \alpha \quad \text{for} \quad \left| e_{ij}/\sigma_{ij} \right| \geq \alpha$$

An α=4 value is suggested (Paatero, 1997) and widely used (Buzcu et al., 2003; Cui et al., 2018; Kim and Hopke, 2004; Li et al., 2004).

The total Q value depends on the size of the data matrix thus, for a more meaningful monitoring of the solution quality Q value is normalized by $Q_{exp}$, representing the degrees of freedom of the model which is both a function of the data matrix size (*n*,*m*) and of the number of factors (*p*).

$$Q_{exp} = n \cdot m - p \cdot (n + m) \sim n \cdot m \quad (5)$$

In the present work, the PMF algorithm was solved using the multilinear engine ME-2 solver (Paatero, 1999). The source Finder toolkit (SoFi Version 6.3 , (Canonaco et al., 2013)) for Igor Pro software package (Wavematrics, Inc., Portland, OR, USA) was used to configure the ME-2 model, for the data downweighting and for post analysis.

Unlike PCA and FA, for the PMF model it is not common to choose the number of factors by means of standardized methods, but it is common practice to investigate solutions with different numbers of factors and observe Q/Qexp, the residual distributions as well as the environmental interpretability of the predicted factors (e.g. Brown et al., 2015; Karanasiou et al., 2011). Choosing the number of factor is a trade-off: increasing the number of factor always reduces the residual but too many factors can lead the model to split a meaningful factor into unrealistic ones (e.g. Yan et al., 2016) obtaining solution which are no more environmentally interpretable. Here several runs with an increasing number of factors from two to seven were done, and the solution was chosen based on the above-mentioned criteria. For the final solution, 50 PMF runs with different seeds where performed and convergence was achieved for each run.

The main sources of uncertainty of a PMF solution are rotational ambiguity and random error in data values (Paatero et al., 2014).

Uncertainty caused by rotational ambiguity arises because bilinear factor analytic models are ill-posed, which means that there are several solutions (G,F) with the same Q value (Henry, 1987). The problem of rotations was discussed in Paatero et al. (2002) and Paatero and Hopke (2009).

To assess rotational ambiguity the method proposed by Paatero et al. (2002) was here implemented. It consists in assessing the rotational ambiguity of the selected PMF solution (base case) by using it as starting point for several runs (i.e. without starting from pseudo-random values) in which different rotational parameter are set. Here, at each run, the factor profiles were constrained around the profiles of the selected base case solution with an α value randomly selected at each iteration between 0 and 1. The α value ($0 \leq \alpha \leq 1$) determines the range in which the resolved factors may deviate from input values.

To assess the total uncertainty of the chosen PMF solution, the effect of random error was investigated through a bootstrap analysis (Davison and Hinkley, 1997). This technique consists in iteratively randomly resampling a subset of data from the original dataset, with their related uncertainty, and performing the PMF analysis on each generated dataset.

At each iteration the data are perturbed by replicating part of the rows of the dataset (i.e. samples) while others are removed so that at each iteration the total amount of samples is kept constant (Paatero et al., 2014). At each bootstrap iteration, the resampled data includes on average ~64% of the total original data per bootstrap run. A total of 100 iterations have been performed. The solution presented in the results session is the average solution calculated on the 100 performed runs. Bootstrap analysis and rotation were implemented using SoFi Pro (Version 6.4).

#### *2.3.3.4 Factor analysis*

To compare the effectiveness of PMF with a more widely used technique, FA analysis has been performed on the same dataset. FA is a multivariate statistical analysis aimed at finding association between parameters by reducing the number of measured parameters to a limited number of factors which can explain the majority of the data variance, avoiding redundant information and noise within the data (Papaioannou et al., 2009). An important aspect of factor analysis is that the resulting factors are orthogonal and therefore not correlated. The software IBM SPSS® was used for performing FA. The correlation matrix was used to perform factor analysis so that no need of a previous data autoscaling was needed (Todeschini, 1998). The Kaiser criterion (Kaiser, 1958) was used to determine the significant factors: it consists in selecting factors with an eigenvalue higher or equal to 1. After the selection of the relevant factors varimax rotation was applied, which is an orthogonal rotation aiming at differentiating the original variables by extracted factor. Factor scores were analysed through GIS system to understand their spatial variability.

### 2.3.4 Results and discussion

#### *2.3.4.1 PMF*

The results of PMF showed that the best number of factors to properly represent the system is five (Table 2). In fact, the first 5 factors were environmentally interpretable, while increasing the number of factors would have only led to split existing factors into two similar sub-factors. Therefore, the solution with five factors was chosen and it is presented here, while the solutions with four and six factors are shown, as an example, respectively in Figure S1 and S2. Results here presented are the average and the standard deviations over the 100 PMF runs.

Qexp is 3780, while for the averaged solution Qrobust is 17605 and Qtrue 30420. Qrobust is the actual error function that is minimized by the PMF model, while Qtrue is the hypothetical Q value calculated, after the solution convergence, without downscaling the scaled residuals outside the (-4 , 4) range.

The EPA PMF User Guide (Norris et al., 2014) reports that similar Qtrue and Qrobust would be a symptom of overestimated analytical uncertainty, implying that the model is not actually running in the robust mode, while the difference between Qtrue and Qrobust is a measure of the presence of sources that are not consistently represented within the dataset. In this case the ratio Between Qtrue and Qrobust is 1.7 which indicates that the considered dataset comes under the second case. Indeed, all the considered variables are positively skewed and particularly 10 out of 14 variables are highly skewed (skewness between 1 and 4). Furthermore ca. 400 values of the data matrix are mathematically considerable outliers. Highly skewed variables and outliers are common in environmental datasets; the risk is that they could have a higher leverage compared to lower values in the PMF solution which is not recommendable, and the role of the robust mode is to reduce these higher leverages so that they can become comparable to those of lower values.

Table 2 shows the relative contribution of each variable to each factor. Namely, it is the F matrix of eqn. 1. The values of Table 2 are related to the concentration of each compound, so that compounds with higher concentration have higher contributions.

**Table 2 - PMF of Factor Profiles**

| | Factor 1 | | Factor 2 | | Factor 3 | | Factor 4 | | Factor 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev | Mean | Std Dev |
| $O_2$ | 0.069 | 0.007 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| P-tot | 0.000 | 0.000 | 0.044 | 0.009 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Alkalinity as $HCO_3$ | 0.431 | 0.030 | 0.170 | 0.079 | 0.098 | 0.037 | 0.761 | 0.002 | 0.451 | 0.023 |
| Cl | 0.022 | 0.003 | 0.044 | 0.017 | 0.019 | 0.008 | 0.005 | 0.001 | 0.040 | 0.004 |
| $NO_3$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.177 | 0.012 |
| $SO_4$ | 0.243 | 0.023 | 0.000 | 0.000 | 0.644 | 0.042 | 0.000 | 0.000 | 0.077 | 0.014 |
| $NH_4$ | 0.000 | 0.000 | 0.267 | 0.040 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Ca | 0.172 | 0.003 | 0.000 | 0.000 | 0.113 | 0.031 | 0.181 | 0.002 | 0.231 | 0.004 |
| Mg | 0.036 | 0.001 | 0.000 | 0.000 | 0.094 | 0.014 | 0.027 | 0.000 | 0.023 | 0.002 |
| Na | 0.019 | 0.001 | 0.290 | 0.058 | 0.001 | 0.001 | 0.023 | 0.001 | 0.001 | 0.000 |
| K | 0.007 | 0.001 | 0.016 | 0.005 | 0.022 | 0.003 | 0.003 | 0.000 | 0.001 | 0.000 |
| As | 0.000 | 0.000 | 0.013 | 0.003 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Fe | 0.000 | 0.000 | 0.148 | 0.026 | 0.002 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Mn | 0.000 | 0.000 | 0.008 | 0.001 | 0.008 | 0.001 | 0.000 | 0.000 | 0.000 | 0.000 |

In order to better appreciate the role of trace elements, which have smaller concentration, Figure 3 shows the fraction of each variable represented by each factor, allowing to understand by which factors every compound is represented.

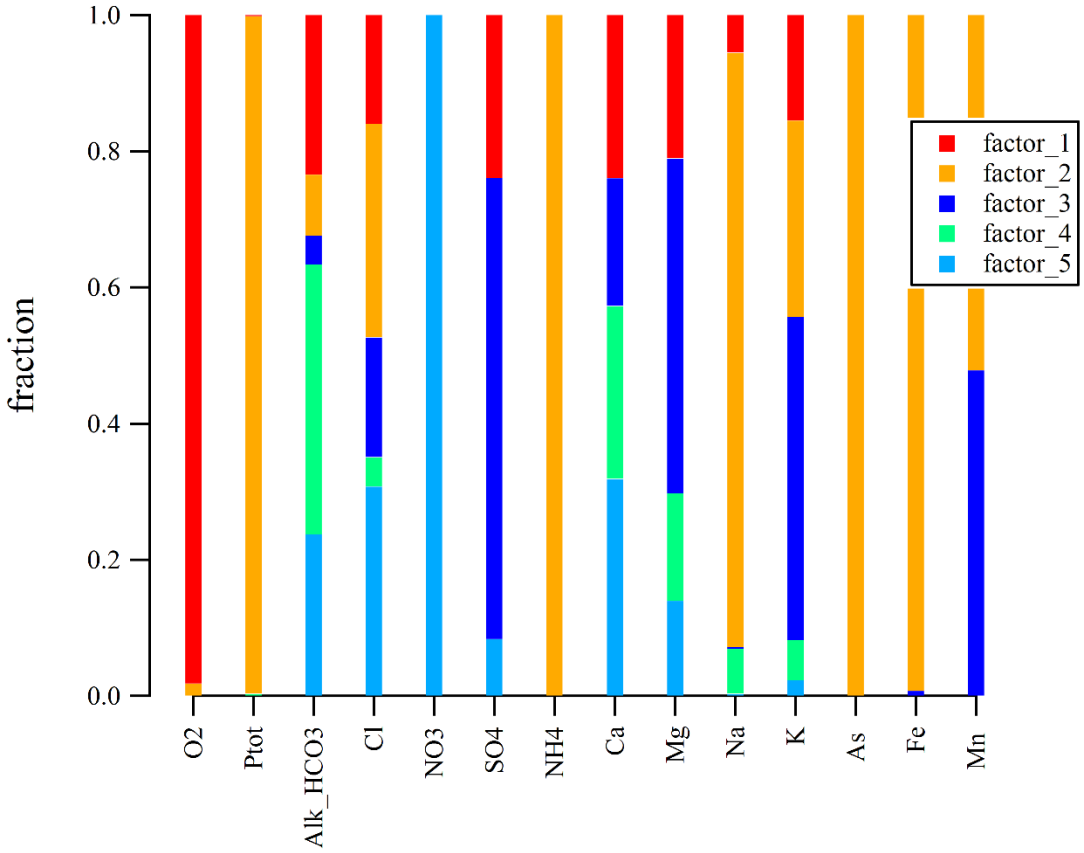



**Figure 3 - Bar chart of relative contribution, as fraction, of the PMF Factors to each variable.**

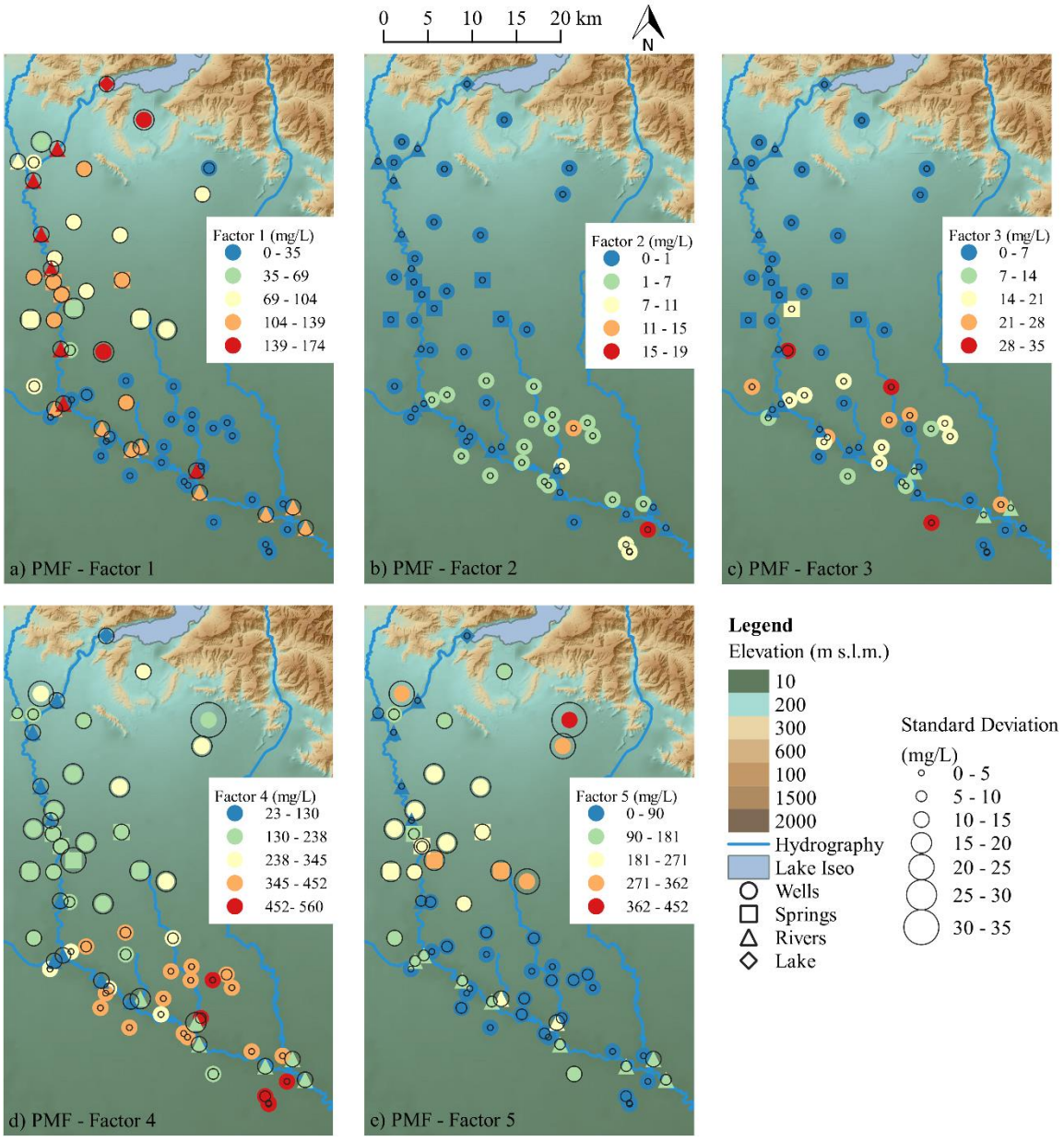



**Figure 4 – Spatial distribution of PMF factor contributions**

The first factor (red bar in Figure 3) is characterized mainly by oxygen and secondly by major ions such as $SO_4$, $HCO_3$, Cl, Ca, Mg, Na, and K. In this factor there is neither nitrate nor reduced species such as As, Fe and Mn. The profile of this factor coincides with the characteristics of surface water bodies directly connected with subalpine lakes. More

specifically, it seems to represent the water of Lake Iseo which has higher $SO_4$ content compared to the other subalpine lakes (Salmaso et al., 2007). The map of the spatial variability of factor 1 (Figure 4a) shows higher values in the Lake Iseo and in the Oglio River (which originates from it) confirming the above. Tributaries have lower values of this factor as compared to the Oglio river, which is due firstly to a lower content of sulphate that in the Oglio river comes mostly from the Lake Iseo and secondarily to a higher anthropogenic impact on the tributaries which lowers the oxygen content. Oxygen in the Oglio River is kept high by the turbulence of the flow, that determines an auto-depuration even if sources of organic matter are present.

Relevant values of factor 1 are also achieved in several wells in the higher plain (Figure 4a). This is related to the aquifer recharge made by irrigation water sourced from the Oglio River, that is the main source of irrigation water in the higher plain (Sect. 2.1). Therefore, this factor, that represents Lake Iseo and Oglio River water, is able to trace aquifer recharge made by irrigation water sourced from Oglio River.

The second factor (orange bar in Figure 3) is characterized by the reduced species As, Fe, $NH_4$ and Mn, together with P-tot and Na. This leads us to consider the second factor as a proxy for reducing conditions. More specifically, it traces reduced conditions that evolved at advanced stages where $NH_4$ is higher as the product of protracted organic matter degradation, $SO_4$ is lower as the result of ongoing or occurred sulfate reduction. Dissolved Fe and Mn are higher as the products of Mn and Fe oxides reduction, that can lead also to higher As and P-tot values (McArthur et al. 2004; Rotiroti et al. 2014). The spatial variability of this factor (Figure 4b) shows that the sampling points with higher values are wells in the lower plain, where As pollution is stronger (i.e. mean As concentrations > 40 ug/L); conversely groundwater samples in the higher plain and surface water samples have no contribution of this factor. On this basis, the second factor, which identifies reducing condition at advanced stages, is able to trace As pollution which affects groundwater in the lower plain.

The third factor (blue bar in Figure 3) is mainly characterized by Mn and $SO_4$ and, secondly, by Mg and K. The higher concentrations of Mn and $SO_4$ indicate the presence of reducing conditions at an earlier stage with respect to factor 2, characterized by the occurrence of Mn-oxide reduction with sulphate reduction that is not favoured yet. Concerning the spatial

distribution of factor 3 (Figure 4c), sampling points with the highest values of this factor are located at the transition between higher and lower plain aquifers and this is consistent with the above since here groundwater is passing from the oxidising conditions of the higher plain to the reducing conditions of the lower plain, so Mn-oxide reduction can be favoured. Conversely groundwater samples in the higher plain and surface water samples have no contribution of this factor.

The fourth factor (green bar in Figure 3) is characterized by only major ions such as $HCO_3$, Cl, Ca, Mg, Na, and K. Major ions concentration in groundwater can be mainly driven by water – rock interactions. Higher major ions concentrations can be reached as groundwater residence times become longer. Therefore, this factor can be associated to groundwater residence times.

Figure 4d shows the spatial variability of this factor highlighting an increase of its value from the higher to the lower plain. This agrees with the hydrogeological features of the study area which has coarse sediments in the higher plain, promoting faster groundwater circulation and so determining shorter residence times, and fine sediments in the lower plain that determine slower groundwater circulation and so longer residence times. Surface water samples downstream show higher values of this factor highlighting the gaining behaviour of the river, and therefore the increasing effect of groundwater to surface water.

The fifth factor (light blue bar in Figure 3) describes the total variability of the variable $NO_3$ with contributions also in terms of Cl, $SO_4$, Ca and Mg. Nitrate occurrence in groundwater can be associated with the use of organic and synthetic fertilizers. Recent studies (Jalali, 2011; Menció et al., 2016) highlighted that an excessive use of fertilizers determines also a linear increase of the major ions Cl, $SO_4$, Ca and Mg in water. Thus, the profile of this fifth factor can represent the impact related to the excessive use of fertilizers in agriculture, not just in terms of $NO_3$ but also considering the contribution of the above-mentioned major ions. This allows to separate the amount of Cl, Ca and Mg determined by this anthropogenic source from the natural amount of these elements in water (factor 4). Concerning the spatial variability of this fifth factor (Figure 4e), it shows high contributions in the samples collected from the higher plain aquifer, that is more vulnerable since it is formed by coarse sediments, and springs, that are a surface outflow of higher plain groundwater. Conversely, in the lower plain the

values of this factor are low due to the occurrence of denitrification that lowers $NO_3$ concentrations close to zero. Concerning the Oglio River, on the other hand, it is possible to notice an enrichment of this factor contribution along its course that attests the increase of $NO_3$ along the river due to the gaining of groundwater from the higher plain which is affected by $NO_3$ (Sect. 2.1).

Figure 5 shows the histogram of the scaled residual of the averaged PMF solution. The distribution is unimodal, centred around zero and symmetric. This indicates that the PMF solution residual are random and that the PMF model is not constantly overestimating or underestimating the observed concentration. The majority of the scaled residual fall within -2 and 2 as suggested by Juntto and Paatero (1994). The peaks at -4 and 4 are due to the robust mode, which substitutes every scaled residual outside the (-4 , 4) range with -4 or 4. Not having scaled residuals at the limit of the robust mode range would lead to equal Qtrue and Qrobust and this would imply too large uncertainties. The amount of data spreading outside this range (with the consequential difference between Qtrue and Qrobust) has been related to inhomogeneous concentrations in the dataset, particularly frequent when working with spatially distributed data (Comero et al., 2009; Norris et al., 2014; Polissar et al., 1998).

The spatial distribution of the residual indicates that the residuals don't have a specific spatial pattern that means that no relevant information has been neglected. The samples collected from the well HL09 (red dot in the northern area in Figure 5b) and from the Mella River (red triangle in the south area, Figure 5b) show the highest residuals. These sampling points located in and around urban environments are more likely affected by urban and industrial sources which are not present within the other samples considered in this study.

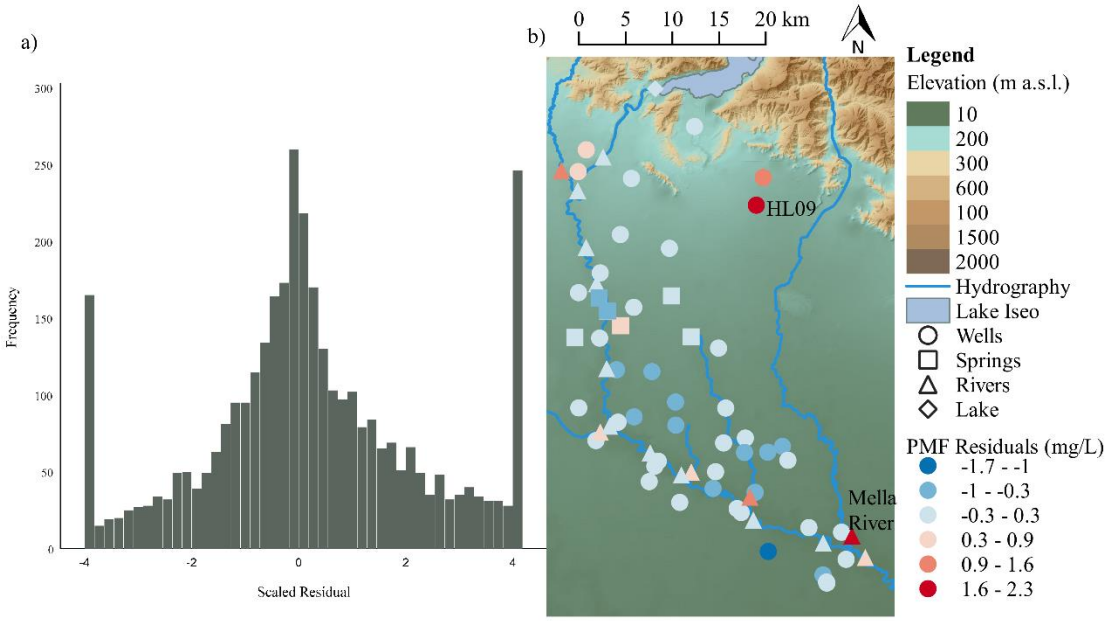



**Figure 5 – a) Scaled residual histogram; b) spatial distribution of the residuals**

### *2.3.4.2 FA*

The results of the FA (table 3 and Figure 6) show that only the first three factors can be considered significant, as they have an eigenvalue higher than 1. These explain the 72.4 % of the variance of the original dataset (cumulative explained variance). These first three factors explain, 37.3%, 24.8% and 10.16% respectively of the total variance. The first factor is mainly represented by P-tot, $NH_4$, Na, As and Fe and it has high negative loadings for the variables $SO_4$. The second one is mainly represented by alkalinity, Cl, $NO_3$, Ca and Mg. The third factor has higher loadings for the variables $O_2$, Cl, $NO_3$, $SO_4$ and K and a strongly negative loading for Mn.

**Table 3 - Factor Loadings of FA**

| Variables | FA Factors | | |
|---|---|---|---|
| | 1 | 2 | 3 |
| $O_2$ (µg/L) | -0.283 | -0.163 | **0.809** |
| P-tot (µg/L) | **0.859** | -0.010 | -0.097 |
| Alkalinity (µg/L) HCO3 | 0.435 | **0.789** | -0.368 |
| Cl (µg/L) | -0.126 | 0.670 | **0.575** |
| $NO_3$ (µg/L) | -0.344 | **0.685** | **0.513** |
| $SO_4$ (µg/L) | **-0.518** | 0.166 | **0.733** |
| $NH_4$ (µg/L) | **0.895** | -0.028 | -0.237 |
| Ca (µg/L) | 0.062 | **0.935** | 0.123 |
| Mg (µg/L) | -0.062 | **0.873** | -0.099 |
| Na (µg/L) | **0.745** | 0.313 | -0.151 |
| K (µg/L) | 0.297 | 0.148 | **0.503** |
| As (µg/L) | **0.664** | -0.138 | -0.009 |
| Fe (µg/L) | **0.756** | -0.007 | -0.313 |
| Mn (µg/L) | 0.374 | 0.065 | **-0.600** |
| Expl Var.% | 37.384 | 24.857 | 10.168 |

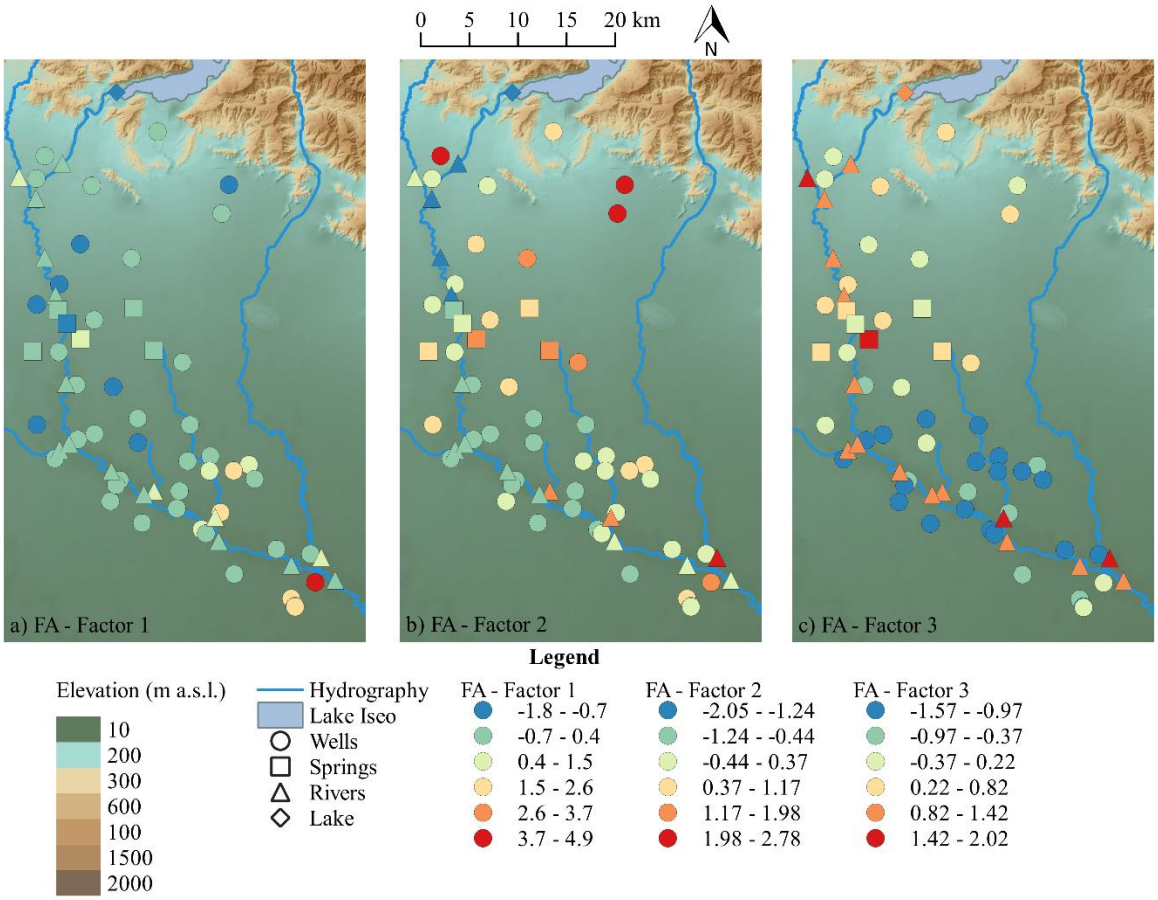



**Figure 6 - FA Factors spatial distribution**

The maps in Figure 6 show the factor scores for the three factors resulting from FA. The first factor has the highest scores in the lower plain in those wells were As pollution is more severe (i.e. mean As > 40 μg/L) that are the same points evidenced also by factor 2 of the PMF. Based on the loadings and the spatial distribution of the scores, it is possible to identify the first factor as a proxy of reducing conditions. Particularly, it seems to include the characteristics of factors 2 and 3 of the PMF but as concerns the spatial distribution, the separation between the higher and the lower plain is less evident. Particularly, it highlights also some differences between the points in the higher plain that cannot be related to As pollution (all these points have As concentrations close to zero) but instead can be related to differences in $SO_4$ and/or Na concentrations.

The second factor, which has higher scores in the higher plain (Figure 6) and a higher loading of $NO_3$ could be associated to anthropogenic impacts mainly related to agricultural activities.

However, high scores of this factor are present also in some wells in the lower plain and that is due to the presence of major ions and alkalinity with high loadings in the factor (Table 3). On this basis, this second factor could be interpreted as a combination of factors 4 and 5 of the PMF.

The third factor resulting from FA represents the highly oxygenated water of surface water bodies. Also in this case it is evident the effect of the irrigation in the higher plain but relatively high scores are evident even within those wells in the lower plain which emerge for highly reduced conditions. This is due to the fact that this factor is also well represented by Mn.

### 2.3.5 Advantages and disadvantages of PMF with respect to FA

Comparing the results of PMF and FA several advantages of the former technique with respect to the latter has emerged. Firstly, it seems that PMF factors depict information in a cleaner way with respect to FA factors, explaining only one main feature or process in each factor. Conversely, the orthogonality constraint in the FA reduces the possibility of discriminating variables which are correlated but related to different processes resulting in factors representing more than one process and/or feature, as described in 3.2, which leads to a more confusing description of the system.

Secondly, PMF considers the uncertainty of the data. Although this requires gathering the information about the uncertainty of the data, it allows to weight differently data with higher uncertainty and to properly work with missing data and data below the detection limit by associating to them a higher uncertainty. FA analysis on the other hand, as conventionally performed, considers all the data with the same weight.

Thirdly, the SoFi tool which was used to perform PMF analysis is suitable for groundwater quality data as it allows to visualize spatial data by importing coordinates and classes vectors related to the samples, which makes more agile the process of selecting and interpreting the solution. An aspect of the PMF that can be considered a disadvantage could be the absence of a predefined method for selecting the number of PMF factors: this choice is based on the comparison of several solutions with different number of factors, as opposite to the FA where well-defined criteria (e.g. Kaiser criterion) are used. This requires a bigger effort when

interpreting the results but, on the other hand, it allows for a higher freedom in exploring the solutions space and a deeper understanding of the data. In the present study PMF allowed to identify more factors than FA, which are representative of environmental phenomena. A second possible disadvantage of PMF is that it works on data expressed as concentration. Therefore, data expressed with different units (e.g. pH, isotopes analysis, Electrical Conductivity and Oxidation Reduction Potential) cannot be directly fed into the PMF model.

A last element for the comparison between PMF and FA that points out how PMF reached a more detailed description of the system is the results of validation with water isotope data. More specifically, water isotopes were used to validate PMF factor 1 and FA factor 3, representing Lake Iseo-Oglio River water and groundwater recharge by irrigation in the higher plain, since Lake Iseo has a more depleted signature (i.e. average value of -9.59 ‰ for $\delta^{18}O$ and -65.12‰ for $\delta^{2}H$) compared to local precipitations (i.e. -7.70‰ for $\delta^{18}O$ and -50.1‰ for $\delta^{2}H$ at the Sarnico station; Longinelli and Selmo, 2003), which constitutes the other relevant source of recharge to groundwater in the area. Figure 7 a and b show the scatterplot of PMF Factor 1 and FA Factor 3 over $\delta^{2}H$ of the wells in the higher plain and Lake Iseo (average over the four field surveys), Figure 7c and d show the scatterplot of $\delta^{18}O$ over $\delta^{2}H$ in which points are coloured respectively by PMF Factor 1 and FA factor 3. Wells in the higher plain which are recharged by irrigation present a more negative isotopic signature, closer to the signature of Lake Iseo. Figure 7a and 7c highlights that PMF factor 1 reproduces more accurately the effect of irrigation on groundwater recharge: samples with a more depleted isotopic signature also have higher contributions of PMF Factor 1 and a linear relationship is evident. On the other hand, FA Factor 3 does not show the same relationship with the isotopic signature of the samples (Figure 7b and 7d). This validation, performed using additional analysis such as isotopes data that were not included in the PMF itself, is an additional evidence that PMF analysis could be a powerful tool to perform exploratory analyses of water quality data to support groundwater quality characterization and the further development of a conceptual model.

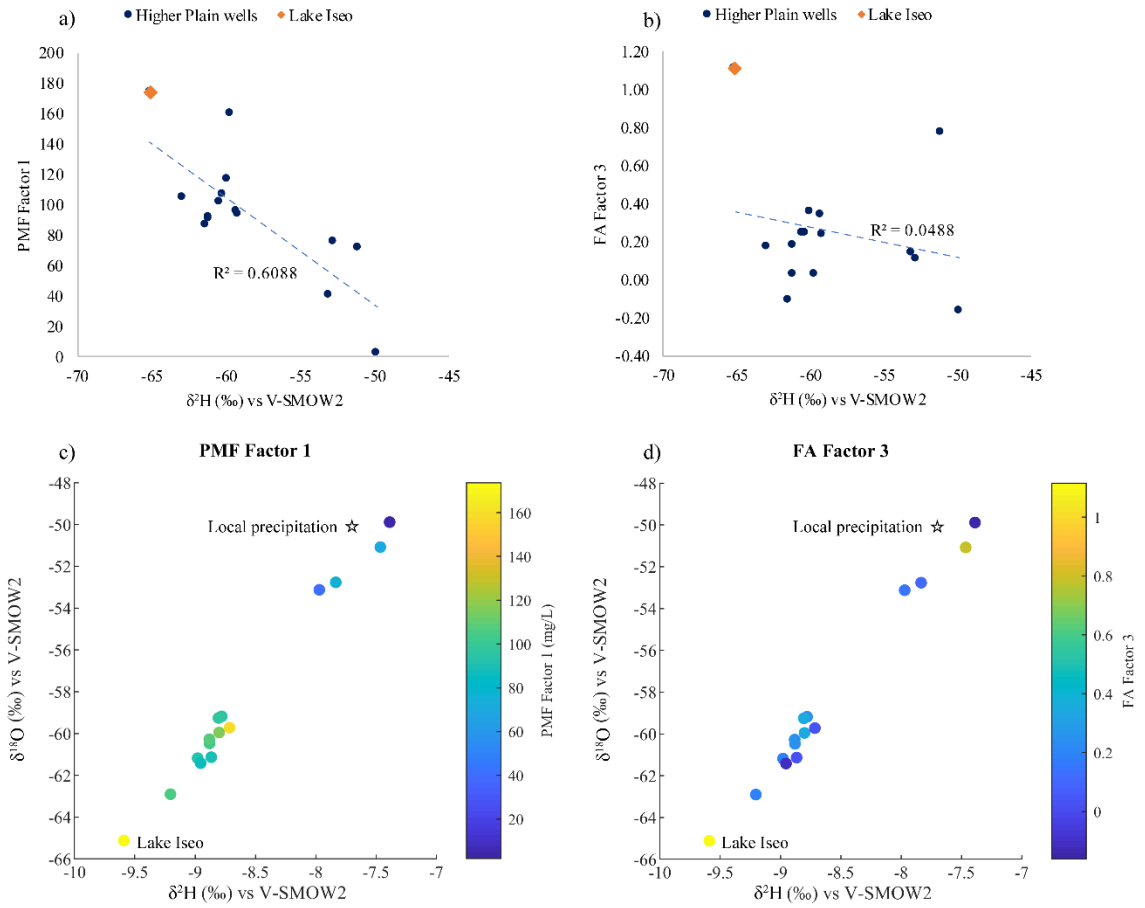



**Figure 7 - a) scatterplot of PMF Factor 1 over $\delta^2H$ ,b) scatterplot of FA Factor 3 over $\delta^2H$, c) scatterplot of $\delta^{18}O$ over $\delta^2H$ coloured by PMF Factor 1 and d) scatterplot of $\delta^{18}O$ over $\delta^2H$ coloured by FA Factor 3.**

### 2.3.6 Conclusions

This work presented a first application of PMF to perform groundwater and surface water quality characterization; PMF results were then compared to those of a FA. The results of both analysis were examined through a GIS approach to appreciate their spatial variability and a validation was performed through the comparison of factor values with water isotope data.

This work proved the ability of PMF to provide an exhaustive characterization of the water quality of the hydro-system under analysis. Indeed, the PMF was able to evidence the main aspects featuring the study area, that are:

- concerning hydrochemical features, the distinction between oxidised groundwater in the higher plain (represented by factor 1) and reduced groundwater in the lower plain (represented by factors 2 and 3);
- concerning pollution phenomena, the presence of the $NO_3$ pollution in the higher plain groundwater and springs (represented by factor 5) and the As pollution in the lower plain groundwater (represented by factor 2);
- concerning main processes governing the system, the occurrence of reducing processes in groundwater in the lower plain, in particular the prevalence of Mn-oxide reduction (represented by factor 3) at the transition between higher and lower plain and the prevalence of sulfate and Fe-oxide reductions (represented by factor 2) in the lower plain; the strong aquifer recharge made by irrigation water in the higher plain (represented by factor 1); the faster groundwater circulation (i.e. shorter residence times) in the higher plain that becomes slower (i.e. longer residence times) in the lower plain (represented by factor 4); the increase in $NO_3$ in the Oglio River after crossing the higher plain due to the gaining of its $NO_3$ affected groundwater (represented by factor 4).

The comparison between PMF and FA highlighted that:

- PMF allows a more detailed description of the system, revealing and characterizing more features and processes compared to FA.
- As opposite to FA, PMF takes into account the analytical uncertainty of the data and can cope with missing and below detection limit data by associating to them a higher uncertainty.
- The positivity constraints of the PMF allow a more environmentally interpretable representation of the system compared to FA.

In summary, PMF could support the identification and quantification of anthropogenic and natural pollution sources, in cases where more than one source is affecting a system. As other data driven techniques the quality of the analysed dataset is crucial, because PMF can only highlight information which is actually gathered within the data. Therefore, PMF can be

applied at any spatial scale as long as the data actually represent the spatial and chemical variability of a system.

Concluding, PMF could be a useful tool to perform water quality characterization in complex hydro-systems where different water resources are involved (e.g. groundwater, lake, rivers and springs) and to support the development of the conceptual model of the system under analysis.

### 2.3.7 Acknowledgements

We thank Valentina Soler of University of Milano – Bicocca for major ions analysis. We thank also Maria Tringali of University of Milano – Bicocca, for the analysis of trace elements and Vittorio Barella and Enrico Allais of ISO4 (Torino, Italy) for performing isotope analysis.

This work was supported by Fondazione Cariplo, grant 2014-1282.

### 2.3.8 References

Alberti, L., Azzellino, A., Colombo, L., Lombi, S., 2016. Cluster Analysis to identify Tetrachloroethylene pollution hotspots for transport numerical model implementation in urban functional area of Milan, Italy, in: 16th International Multidisciplinary Scientific Conference SGEM2016, Book 1. https://doi.org/10.5593/SGEM2016/B11/S02.091

Belis, C.A., Cancelinha, J., Duane, M., Forcina, V., Pedroni, V., Passarella, R., Tanet, G., Douglas, K., Piazzalunga, A., Bolzacchini, E., Sangiorgi, G., Perrone, M.G., Ferrero, L., Fermo, P., Larsen, B.R., 2011. Sources for PM air pollution in the Po Plain, Italy: I. Critical comparison of methods for estimating biomass burning contributions to benzo(a)pyrene. Atmos. Environ. 45, 7266–7275. https://doi.org/10.1016/j.atmosenv.2011.08.061

Blake, S., Henry, T., Murray, J., Flood, R., Muller, M.R., Jones, A.G., Rath, V., 2016. Applied Geochemistry Compositional multivariate statistical analysis of thermal groundwater provenance : A hydrogeochemical case study from Ireland. Appl. Geochemistry 75, 171–188. https://doi.org/10.1016/j.apgeochem.2016.05.008

Bozzetti, C., Sosedova, Y., Xiao, M., Daellenbach, K.R., Ulevicius, V., Dudoitis, V., Mordas, G., Byčenkiene, S., Plauškaite, K., Vlachou, A., Golly, B., Chazeau, B., Besombes, J.L., Baltensperger, U., Jaffrezo, J.L., Slowik, J.G., El Haddad, I., Prévôt, A.S.H., 2017.

Argon offline-AMS source apportionment of organic aerosol over yearly cycles for an urban, rural, and marine site in northern Europe. Atmos. Chem. Phys. 17, 117–141. https://doi.org/10.5194/acp-17-117-2017

Brown, S.G., Eberly, S., Paatero, P., Norris, G.A., 2015. Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. Sci. Total Environ. 518–519, 626–635. https://doi.org/10.1016/j.scitotenv.2015.01.022

Buzcu, B., Fraser, M.P., Kulkarni, P., Chellam, S., 2003. Source identification and apportionment of fine particulate matter in Houston, TX, using positive matrix factorization. Environ. Eng. Sci. 20, 533–545.

Canonaco, F., Crippa, M., Slowik, J.G., Baltensperger, U., Prévôt, A.S.H., 2013. SoFi, an IGOR-based interface for the efficient use of the generalized multilinear engine (ME-2) for the source apportionment: ME-2 application to aerosol mass spectrometer data. Atmos. Meas. Tech. 6, 3649–3661. https://doi.org/10.5194/amt-6-3649-2013

Carraro, A., Fabbri, P., Giaretta, A., Peruzzoa, L., Tateo, F., Tellini, F., 2015. Effects of redox conditions on the control of arsenic mobility in shallow alluvial aquifers on the Venetian Plain (italy). Sci. Total Environ. 532, 581–594. https://doi.org/10.1016/j.scitotenv.2015.06.003

Cinnirella, S., Buttafuoco, G., Pirrone, N., 2005. Stochastic analysis to assess the spatial distribution of groundwater nitrate concentrations in the Po catchment (Italy). Environ. Pollut. 133, 569–580. https://doi.org/10.1016/j.envpol.2004.06.020

Comero, S., Capitani, L., Gawlik, B.M., 2009. Positive Matrix Factorisation (PMF) An introduction to the chemometric evaluation of environmental monitoring data using PMF. https://doi.org/10.2788/2497

Comero, S., Locoro, G., Free, G., Vaccaro, S., De Capitani, L., Gawlik, B.M., 2011. Characterisation of Alpine lake sediments using multivariate statistical techniques. Chemom. Intell. Lab. Syst. 107, 24–30. https://doi.org/10.1016/j.chemolab.2011.01.002

Comero, S., Servida, D., De Capitani, L., Gawlik, B.M., 2012. Geochemical characterization of an abandoned mine site: A combined positive matrix factorization and GIS

approach compared with principal component analysis. J. Geochemical Explor. 118, 30–37. https://doi.org/10.1016/j.gexplo.2012.04.003

Cui, Y.Y., Liu, S., Bai, Z., Bian, J., Li, D., Fan, K., McKeen, S.A., Watts, L.A., Ciciora, S.J., Gao, R.-S., 2018. Religious burning as a potential major source of atmospheric fine aerosols in summertime Lhasa on the Tibetan Plateau. Atmos. Environ. 181, 186–191. https://doi.org/https://doi.org/10.1016/j.atmosenv.2018.03.025

Daellenbach, K.R., Stefenelli, G., Bozzetti, C., Vlachou, A., Fermo, P., Gonzalez, R., Piazzalunga, A., Colombi, C., Canonaco, F., Hueglin, C., Kasper-Giebl, A., Jaffrezo, J.-L., Bianchi, F., Slowik, J.G., Baltensperger, U., El Haddad, I., Prévôt, A.S.H., 2017. Long-term chemical analysis and organic aerosol source apportionment at 9 sites in Central Europe: Source identification and uncertainty assessment. Atmos. Chem. Phys. Discuss. 1–36. https://doi.org/10.5194/acp-2017-124

Davison, A.C., Hinkley, D.V., 1997. Bootstrap methods and their application. Cambridge university press.

Devic, G., Djordjevic, D., Sakan, S., 2014. Natural and anthropogenic factors affecting the groundwater quality in Serbia. Sci. Total Environ. 468–469, 933–942. https://doi.org/10.1016/j.scitotenv.2013.09.011

Fortner, E., Onasch, T., Canagaratna, M., Williams, L.R., Lee, T., Jayne, J., Worsnop, D., 2018. Examining the chemical composition of black carbon particles from biomass burning with SP-AMS. J. Aerosol Sci. 120, 12–21. https://doi.org/https://doi.org/10.1016/j.jaerosci.2018.03.001

Giuliano, G., 1995. Ground water in the Po basin: some problems relating to its use and protection. Sci Total Environ 171 (1):17-27. doi:https://doi.org/10.1016/0048-9697(95)04682-1

Grøn, C., Hansen, J.B., Magnusson, B., Nordbotten, A., Krysell, M., Andersen, Kirsten Jebjerg Lund, U., 2007. Uncertainty from sampling - a Nordtest handbook for sampling planners on sampling quality assurance and uncertainty estimation.

Gu, H., Ma, F., Guo, J., Zhao, H., Lu, R., Liu, G., 2018. A Spatial Mixing Model to Assess Groundwater Dynamics Affected by Mining in a Coastal Fractured Aquifer, China. Mine Water Environ. 37, 405–420. https://doi.org/10.1007/s10230-017-0505-x

Haji Gholizadeh, M., Melesse, A.M., Reddi, L., 2016. Water quality assessment and apportionment of pollution sources using APCS-MLR and PMF receptor modeling techniques in three major rivers of South Florida. Sci. Total Environ. 566–567, 1552–1567. https://doi.org/10.1016/j.scitotenv.2016.06.046

Henry, R.C., 1987. Current factor analysis receptor models are ill-posed. Atmos. Environ. 21, 1815–1820.

Hopke, P.K., 2000. A guide to positive matrix factorization. Work. UNMIX PMF as Appl. to PM2 5, 600.

Jalali, M., 2011. Nitrate pollution of groundwater in Toyserkan, western Iran. Environ. Earth Sci. 62, 907–913. https://doi.org/10.1007/s12665-010-0576-5

Juntto, S., Paatero, P., 1994. Analysis of daily precipitation data by positive matrix factorization. Environmetrics 5, 127–144.

Karanasiou, A., Moreno, T., Amato, F., Lumbreras, J., Narros, A., Borge, R., Tobías, A., Boldo, E., Linares, C., Pey, J., Reche, C., Alastuey, A., Querol, X., 2011. Road dust contribution to PM levels - Evaluation of the effectiveness of street washing activities by means of Positive Matrix Factorization. Atmos. Environ. 45, 2193–2201. https://doi.org/10.1016/j.atmosenv.2011.01.067

Kim, E., Hopke, P.K., 2004. Source apportionment of fine particles in Washington, DC, utilizing temperature-resolved carbon fractions. J. Air Waste Manage. Assoc. 54, 773–785.

Koh, D.C., Chae, G.T., Ryu, J.S., Lee, S.G., Ko, K.S., 2016. Occurrence and mobility of major and trace elements in groundwater from pristine volcanic aquifers in Jeju Island, Korea. Appl. Geochemistry 65, 87–102. https://doi.org/10.1016/j.apgeochem.2015.11.004

Kuang, B.Y., Lin, P., Huang, X.H.H., Yu, J.Z., 2015. Sources of humic-like substances in the Pearl River Delta, China: Positive matrix factorization analysis of PM2.5 major components and source markers. Atmos. Chem. Phys. 15, 1995–2008. https://doi.org/10.5194/acp-15-1995-2015

Lasagna, M., De Luca, D.A., Franchino, E., 2016. Nitrate contamination of groundwater in the western Po Plain (Italy): the effects of groundwater and surface water interactions. Environ. Earth Sci. 75, 1–16. https://doi.org/10.1007/s12665-015-5039-6

Leoni, B., Garibaldi, L., Gulati, R.D., 2014a. How does interannual trophic variability caused by vertical water mixing affect reproduction and population density of the Daphnia longispina group in Lake Iseo, a deep stratified lake in Italy? Inl. Waters. https://doi.org/10.5268/IW-4.2.663

Leoni, B., Marti, C.L., Imberger, J., Garibaldi, L., 2014b. Summer spatial variations in phytoplankton composition and biomass in surface waters of a warm-temperate, deep, oligo-holomictic lake: Lake Iseo, Italy. Inl. Waters. https://doi.org/10.5268/IW-4.3.569

Li, Z., Hopke, P.K., Husain, L., Qureshi, S., Dutkiewicz, V.A., Schwab, J.J., Drewnick, F., Demerjian, K.L., 2004. Sources of fine particle composition in New York city. Atmos. Environ. 38, 6521–6529.

Longinelli, A., Selmo, E., 2003. Isotopic composition of precipitation in Italy: a first overall map. J. Hydrol. 270, 75–88. https://doi.org/https://doi.org/10.1016/S0022-1694(02)00281-0

Marti, C.L., Imberger, J., Garibaldi, L., Leoni, B., 2016. Using time scales to characterize phytoplankton assemblages in a deep subalpine lake during the thermal stratification period: Lake Iseo, Italy. Water Resour. Res. https://doi.org/10.1002/2015WR017555

Martinelli, G., Dadomo, A., Luca, D.A. De, Mazzola, M., Lasagna, M., Pennisi, M., Pilla, G., Sacchi, E., Saccon, P., 2018. Nitrate sources , accumulation and reduction in groundwater from Northern Italy : Insights provided by a nitrate and boron isotopic database Applied Geochemistry Nitrate sources , accumulation and reduction in groundwater from Northern Italy : Insights pr. Appl. Geochemistry 91, 23–35. https://doi.org/10.1016/j.apgeochem.2018.01.011

Mcarthur, J.M., Banerjee, D.M., Hudson-edwards, K.A., Mishra, R., Talukder, T., Lowry, D., Houghton, S., Chadha, D.K., 2004. Natural organic matter in sedimentary basins and its relation to arsenic in anoxic ground water : the example of West Bengal and its worldwide implications 19, 1255–1293. https://doi.org/10.1016/j.apgeochem.2004.02.001

Menció, A., Mas-Pla, J., Otero, N., Regàs, O., Boy-Roura, M., Puig, R., Bach, J., Domènech, C., Zamorano, M., Brusi, D., Folch, A., 2016. Nitrate pollution of groundwater; all

right. . ., but nothing else? Sci. Total Environ. 539, 241–251. https://doi.org/10.1016/j.scitotenv.2015.08.151

Mohr, C., DeCarlo, P.F., Heringa, M.F., Chirico, R., Richter, R., Crippa, M., Querol, X., Baltensperger, U., Prévôt, A.S.H., 2015. Spatial Variation of Aerosol Chemical Composition and Organic Components Identified by Positive Matrix Factorization in the Barcelona Region. Environ. Sci. Technol. 49, 10421–10430. https://doi.org/10.1021/acs.est.5b02149

Mohr, C., DeCarlo, P.F., Heringa, M.F., Chirico, R., Slowik, J.G., Richter, R., Reche, C., Alastuey, A., Querol, X., Seco, R., Peñuelas, J., Jiḿenez, J.L., Crippa, M., Zimmermann, R., Baltensperger, U., Prévôt, A.S.H., 2012. Identification and quantification of organic aerosol from cooking and other sources in Barcelona using aerosol mass spectrometer data. Atmos. Chem. Phys. 12, 1649–1665. https://doi.org/10.5194/acp-12-1649-2012

Molinari, A., Guadagnini, L., Marcaccio, M., Guadagnini, A., 2012. Natural background levels and threshold values of chemical species in three large-scale groundwater bodies in Northern Italy. Sci. Total Environ. 425, 9–19. https://doi.org/10.1016/j.scitotenv.2012.03.015

Norris, G.A., Duvall, R., Brown, S., Bai, S., 2014. EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide

Paatero, P., 1999. The multilinear engine—a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. J. Comput. Graph. Stat. 8, 854–888.

Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. Chemom. Intell. Lab. Syst. 37, 23–35.

Paatero, P., Eberly, S., Brown, S.G., Norris, G.A., 2014. Methods for estimating uncertainty in factor analytic solutions 781–797. https://doi.org/10.5194/amt-7-781-2014

Paatero, P., Hopke, P.K., 2009. Rotational Tools for Factor Analytic Models 91–100. https://doi.org/10.1002/cem.1197

Paatero, P., Hopke, P.K., 2003. Discarding or downweighting high-noise variables in factor analytic models. Anal. Chim. Acta 490, 277–289. https://doi.org/10.1016/S0003-2670(02)01643-4

Paatero, P., Hopke, P.K., Song, X., Ramadan, Z., 2002. Understanding and controlling rotations in factor analytic models 60, 253–264.

Paatero, P., Tapper, U., 1994. Positive Matrix Factorization: A Non-Negative Factor Model with Optimal Utilization of Error Estimates of Data Values. Environmetrics 5, 111–126.

Papaioannou, A., Mavridou, A., Hadjichristodoulou, C., Papastergiou, P., Pappa, O., Dovriki, E., Rigas, I., 2009. Application of multivariate statistical methods for groundwater physicochemical and biological quality assessment in the context of public health. Environ. Monit. Assess. 170, 87–97. https://doi.org/10.1007/s10661-009-1217-x

Pareeth, S., Bresciani, M., Buzzi, F., Leoni, B., Lepori, F., Ludovisi, A., Morabito, G., Adrian, R., Neteler, M., Salmaso, N., 2017. Warming trends of perialpine lakes from homogenised time series of historical satellite and in-situ data. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2016.10.199

Pekey, H., Doğan, G., 2013. Application of positive matrix factorisation for the source apportionment of heavy metals in sediments: A comparison with a previous factor analysis study. Microchem. J. 106, 233–237. https://doi.org/10.1016/j.microc.2012.07.007

Perego, R., Bonomi, T., Fumagalli, L., Benastini, V., Aghib, F., Rotiroti, M., Cavallin, A., 2014. 3D reconstruction of the multi-layer aquifer in a Po Plain area. Rend Online Soc Geol It 30:41-44. doi:10.3301/ROL.2014.09

Phung, D., Huang, C., Rutherford, S., Dwirahmadi, F., Chu, C., Wang, X., Nguyen, M., Nguyen, N.H., Do, C.M., Nguyen, T.H., Dinh, T.A.D., 2015. Temporal and spatial assessment of river surface water quality using multivariate statistical techniques: a study in Can Tho City, a Mekong Delta area, Vietnam. Environ. Monit. Assess. 187. https://doi.org/10.1007/s10661-015-4474-x

Polissar, A. V, Hopke, P.K., Paatero, P., Malm, W.C., Sisler, J.F., 1998. Atmospheric aerosol over Alaska: 2. Elemental composition and sources. J. Geophys. Res. Atmos. 103, 19045–19057.

Reff, A., Eberly, S.I., Bhave, P. V., 2007. Receptor Modeling of Ambient Particulate Matter Data Using Positive Matrix Factorization: Review of Existing Methods. J. Air Waste Manage. Assoc. 57, 146–154. https://doi.org/10.1080/10473289.2007.10465319

Rogora, M., Buzzi, F., Dresti, C., Leoni, B., Lepori, F., Mosello, R., Patelli, M., Salmaso, N., 2018. Climatic effects on vertical mixing and deep-water oxygen content in the subalpine lakes in Italy. Hydrobiologia. https://doi.org/10.1007/s10750-018-3623-y

Rotiroti M, Bonomi T, Sacchi E, McArthur JM, Stefania GA, Zanotti C, Taviani S, Patelli M, Nava V, Soler V, Fumagalli L, Leoni B., 2019a. The effects of irrigation on groundwater quality and quantity in a human-modified hydro-system: the Oglio River basin, Po Plain, northern Italy. Sci Total Environ, https://doi.org/10.1016/j.scitotenv.2019.03.427.

Rotiroti M, Jakobsen R, Fumagalli L, Bonomi T., 2015. Arsenic release and attenuation in a multilayer aquifer in the Po Plain (northern Italy): Reactive transport modeling. Appl Geochem 63:599-609. doi:http://dx.doi.org/10.1016/j.apgeochem.2015.07.001

Rotiroti, M., Sacchi, E., Fumagalli, L., Bonomi, T., 2014. Origin of arsenic in groundwater from the multilayer aquifer in cremona (Northern Italy). Environ. Sci. Technol. 48 (10):, 5395–5403. https://doi.org/10.1021/es405805v

Rotiroti M, McArthur J, Fumagalli L, Stefania GA, Sacchi E, Bonomi T., 2017. Pollutant sources in an arsenic-affected multilayer aquifer in the Po Plain of Italy: Implications for drinking-water supply. Sci Total Environ 578:502-512. doi:http://dx.doi.org/10.1016/j.scitotenv.2016.10.215

Rotiroti M, Zanotti C, Fumagalli L, Taviani S, Stefania GA, Patelli M, Nava V, Soler V, Sacchi E and Leoni B., 2019b. Multivariate statistical analysis supporting the hydrochemical characterization of groundwater and surface water: a case study in northern Italy. Rend Online Soc Geol It, 47:90-96. doi:https://doi.org/10.3301/ROL.2019.17

Sacchi E, Acutis M, Bartoli M, Brenna S, Delconte CA, Laini A, Pennisi M., 2013. Origin and fate of nitrates in groundwater from the central Po plain: Insights from isotopic investigations. Appl Geochem 34:164-180. doi:http://dx.doi.org/10.1016/j.apgeochem.2013.03.008

Salmaso, N., Morabito, G., Garibaldi, L., Mosello, R., 2007. Trophic development of the deep lakes south of the Alps: a comparative analysis. Fundam. Appl. Limnol. / Arch. für Hydrobiol. 170, 177–196. https://doi.org/10.1127/1863-9135/2007/0170-0177

Shao, Y., Wang, Y., Xu, X., Wu, X., Jiang, Z., He, S., Qian, K., 2014. Occurrence and source apportionment of PAHs in highly vulnerable karst system. Sci. Total Environ. 490, 153–160. https://doi.org/10.1016/j.scitotenv.2014.04.128

Shrestha, S.M., Rijal, K., Pokhrel, M.R., 2016. Assessment of Heavy Metals in Deep Groundwater Resources of the Kathmandu. J. Environ. Prot. (Irvine,. Calif). 7, 516–531.

Soonthornnonda, P., Christensen, E.R., 2008. Source apportionment of pollutants and flows of combined sewer wastewater. Water Res. 42, 1989–1998. https://doi.org/10.1016/j.watres.2007.11.034

Sowlat, M.H., Hasheminassab, S., Sioutas, C., 2016. Source apportionment of ambient particle number concentrations in central Los Angeles using positive matrix factorization (PMF). Atmos. Chem. Phys. 16, 4849–4866. https://doi.org/10.5194/acp-16-4849-2016

Stefania, G.A., Zanotti, C., Bonomi, T., Fumagalli, L., Rotiroti, M., 2018. Determination of trigger levels for groundwater quality in landfills located in historically human-impacted areas. Waste Manag. 75, 400–406. https://doi.org/10.1016/j.wasman.2018.01.043

Todeschini, R., 1998. Introduzione alla chemiometria. Edises, Napoli.

Visser, S., Slowik, J.G., Furger, M., Zotter, P., Bukowiecki, N., Canonaco, F., Flechsig, U., Appel, K., Green, D.C., Tremper, A.H., Young, D.E., Williams, P.I., Allan, J.D., Coe, H., Williams, L.R., Mohr, C., Xu, L., Ng, N.L., Nemitz, E., Barlow, J.F., Halios, C.H., Fleming, Z.L., Baltensperger, U., Prévôt, A.S.H., 2015. Advanced source apportionment of size-resolved trace elements at multiple sites in London during winter. Atmos. Chem. Phys. 15, 11291–11309. https://doi.org/10.5194/acp-15-11291-2015

Witczak, S., 2006. BRIDGE - Summary guidance and recommendations on Sampling, Measuring and Quality Assurance.

Yan, C., Nie, W., Äijälä, M., Rissanen, M.P., Canagaratna, M.R., Massoli, P., Junninen, H., Jokinen, T., Sarnela, N., Häme, S.A.K., Schobesberger, S., Canonaco, F., Yao, L., Prévôt, A.S.H., Petäjä, T., Kulmala, M., Sipilä, M., Worsnop, D.R., Ehn, M., 2016. Source characterization of highly oxidized multifunctional compounds in a boreal

forest environment using positive matrix factorization. Atmos. Chem. Phys. 16, 12715–12731. https://doi.org/10.5194/acp-16-12715-2016

Zheng, T., Ran, Y., Chen, L., 2014. Polycyclic aromatic hydrocarbons (PAHs) in rural soils of Dongjiang River Basin: Occurrence, source apportionment, and potential human health risk. J. Soils Sediments 14, 110–120. https://doi.org/10.1007/s11368-013-0753-8

### 2.3.9 Supplementary materials

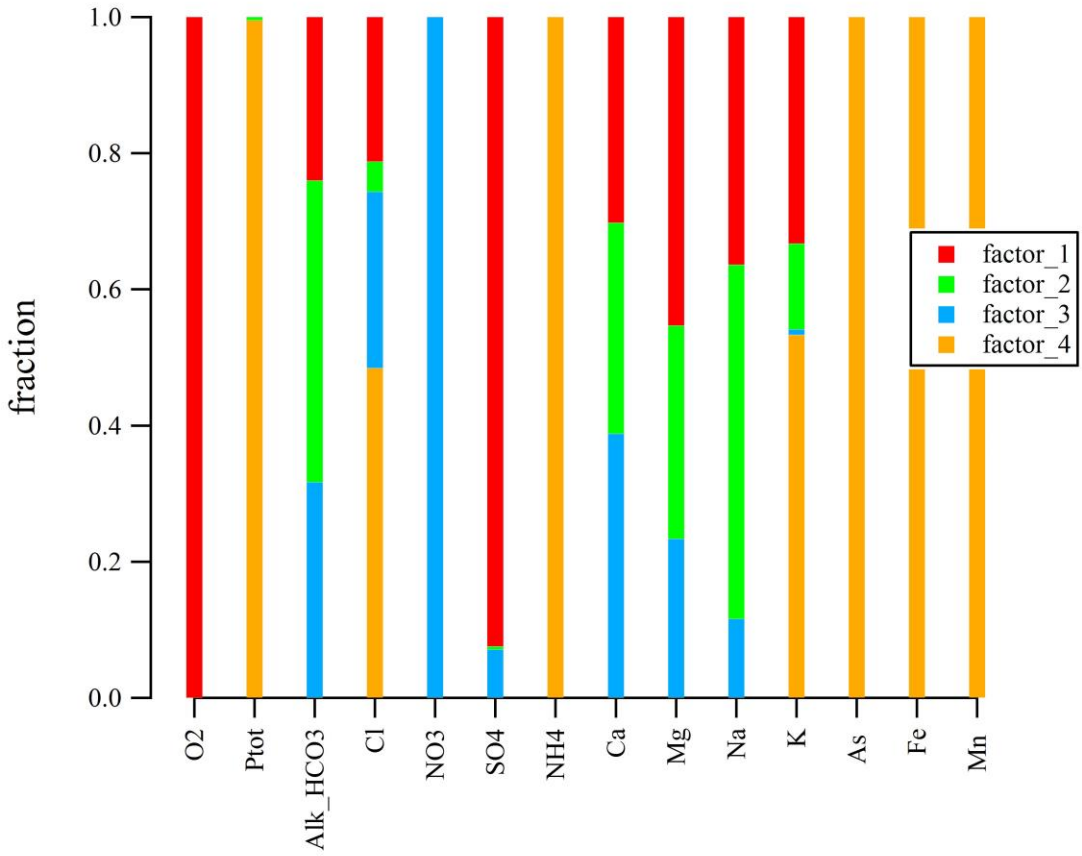



**Figure S1 - Bar chart of relative contribution, as fraction, of the PMF factors to each variable for the PMF solution with 4 factors**

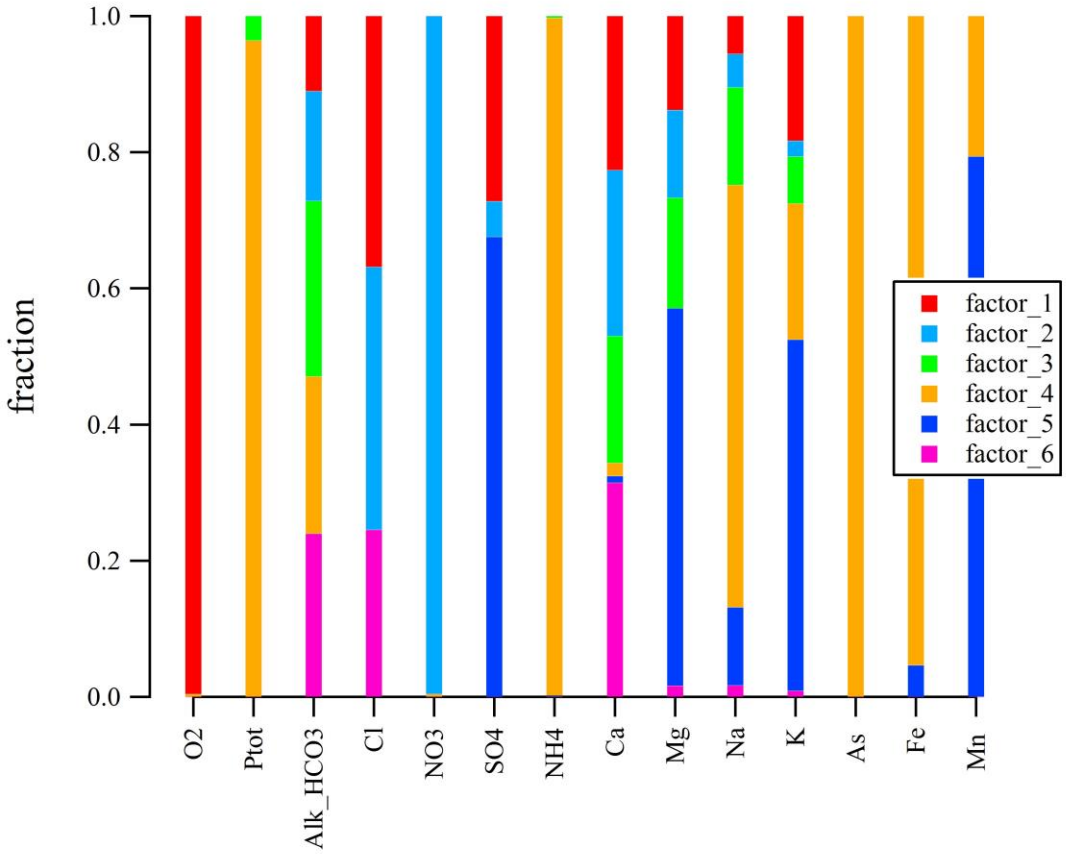



**Figure S2 - Bar chart of relative contribution, as fraction, of the PMF factors to each variable for the PMF solution with 6 factors**

# Chapter 3
# Groundwater quantity

## 3.1 Introduction

A large part of the Mediterranean countries water supply is provided by karst aquifers, nevertheless there is only limited knowledge about their sensitivity to climatic change (Hartmann et al., 2014). Climate change simulations point out that in the next 90 years Mediterranean regions will be exposed to lower precipitation and higher temperature (Christensen et al., 2007). Furthermore, an increase in terms of hydroclimatic extremes can be expected in the Mediterranean areas in terms of both, hydrological floods (Hirabayashi et al., 2013) and droughts (Milly et al., 2005). It is important but, at the same time, difficult to assess the impact of these changes on karst water resources (Hartmann et al., 2014). A first step for a wise groundwater resource management is a proper attitude and vision of the plausible future events which may happen (Choubin and Malekian, 2017), in this scope developing forecasting tools for groundwater level becomes a crucial aspect of groundwater management.

Typically, numerical flow models are used for simulating a groundwater flow system and predicting the groundwater level fluctuation. These models determine a governing equation, by simplifying the physics of the flow system and solve it with numerical methods given proper initial and boundary conditions (Spitz et al., 1996).

For this approach, to reach reliable predictions, a large amount of precise information is required to reproduce physical properties of the domain and the model parameters and to calibrate the model simulations which are sometimes difficult to obtain (Yoon et al., 2011).

This task addresses the problem of groundwater level forecasting in a karst aquifer based on previous groundwater levels and meteorological data using databased time series models. These models require only time series of groundwater levels and meteorological data but the forecast is limited to temporal variations at a fixed location.

This study involves the karst thermal aquifer system of Monsummano Terme in central Italy (PT). This system is located at the transition between the northern Apennines and the Arno Plain, consisting in a fractured carbonate aquifer. A previous study reports that the Monsumanno Terme aquifer has a short extension (~1-2 km of length and 600-700 m of thickness) and that it can be considered a closed system being independent and physically separated from the nearby Montecatini Terme aquifer system and that it has no interrelations with surface water bodies. The system is recharged upstream by local precipitation in the closeby hilly area of carbonates outcrops (Monsummano Alto, approximately 150-300 m a.s.l) and discharges downstream through natural springs and well abstractions (Monsummano Terme, at ~50 m a.s.l.). Moreover, a linear and quick response of the groundwater levels in the discharge area to the precipitation in the recharge area has been identified in previous studies (Grassi et al., 2011).

In this task of the present PhD work, data analysis of the time series of groundwater level and meteorological data was performed in order to investigate the effectiveness of data driven techniques in developing forecasting tool. Particularly the reliability of the developed forecasting tool was tested in the scope of hydroclimatic extremes. This task was performed in two successive phases:

1) The first phase, entitled "Groundwater level forecasting using linear time series modeling: the case study of the thermal aquifer system of Monsummano Terme (central Italy)" consists in the exploratory analysis of the groundwater level and precipitation time series aimed at identifying the information contained within the available data, in order to steer the models implementation. Since the studied data consist in time-domain sequences, the exploratory analysis was conducted by means of techniques belonging to the family of the time-series analysis. More specifically, Auto Correlation and Partial Autocorrelation were applied as univariate statistics, whereas Cross Correlation and Impulse Response were used as bivariate statistics. The

point of these analysis is to investigate the "memory effect" of the system, determining how much the groundwater level at a certain time t is affected by previous values of groundwater level (Auto Correlation and Partial Autocorrelation) and past precipitation events (Cross Correlation and Impulse Response). The results of this exploratory analysis were used as a base for the development of the forecasting tools in the second phase of the work.

2) The second phase, entitled "Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system" consists in the development of forecasting models. The goal of this work is to compare the effectiveness of a linear Autoregressive model with exogenous input (ARx) with Neural Network – ARx combined models (NNARx) on short and long term forecasting in a groundwater system where a linear input-output relationship can be assumed on the basis of the conceptual model.

   Neural Networks models are often described as "black box" models capable of providing relatively accurate forecasts thanks to their ability to adapt to dynamic, and non-linear changes (Adamowski and Chan, 2011; Daliakopoulos et al., 2005; Khalil et al., 2015). On the other hand, methods for developing neural networks models for hydrological problems are not yet well established (Maier et al., 2010) and there is a progressive advance towards improving the performance of neural networks models (Sudheer and Kasiviswanathan, 2017). This work specifically focuses on the reliability of these models as forecasting tools for hydrological extremes by testing them on data which exceed the range of the data used for the models' training.

**References**

Adamowski, J., Chan, H.F., 2011. A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol. 407, 28–40. https://doi.org/10.1016/j.jhydrol.2011.06.013

Choubin, B., Malekian, A., 2017. Combined gamma and M-test-based ANN and ARIMA models for groundwater fluctuation forecasting in semiarid regions. Environ. Earth Sci. 76, 1–10. https://doi.org/10.1007/s12665-017-6870-8

Christensen, J.H., Hewitson, B., Busuioc, A., Chen, A., Gao, X., Held, R., Jones, R., Kolli, R.K., Kwon, W.K., Laprise, R., others, 2007. Regional climate projections, in: Climate Change, 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, University Press, Cambridge, Chapter 11. pp. 847–940.

Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K., 2005. Groundwater level forecasting using artificial neural networks. J. Hydrol. 309, 229–240. https://doi.org/10.1016/j.jhydrol.2004.12.001

Grassi, S., Doveri, M., Ellero, A., Palmieri, F., Vaselli, L., Calvi, E., Trifirò, S., 2011. Studio dei sistemi termali di Montecatini e Monsummano Terme.

Hartmann, A., Mudarra, M., Andreo, B., Marin, A., Wagener, T., Lange, J., 2014. Modeling spatiotemporal impacts of hydroclimatic extremes on groundwater recharge at a Mediterranean karst aquifer. Water Resour. Res. 50, 6507–6521. https://doi.org/10.1002/2014WR015685.Received

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., Kim, H., Kanae, S., 2013. Global flood risk under climate change. Nat. Clim. Chang. 3, 816.

Khalil, B., Broda, S., Adamowski, J., Ozga-Zielinski, B., Donohoe, A., 2015. Short-term forecasting of groundwater levels under conditions of mine-tailings recharge using wavelet ensemble neural network models. Hydrogeol. J. 23, 121–141. https://doi.org/10.1007/s10040-014-1204-3

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 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

Milly, P.C.D., Dunne, K.A., Vecchia, A. V, 2005. Global pattern of trends in streamflow and water availability in a changing climate. Nature 438, 347–350. https://doi.org/10.1038/nature04312

Spitz, K., Moreno, J., others, 1996. A practical guide to groundwater and solute transport modeling. John Wiley and sons.

Sudheer, K.S., Kasiviswanathan, K.P., 2017. Methods used for quantifying the prediction uncertainty of artificial neural network based hydrologic models. Stoch. Environ. Res. Risk Assess. 31, 1659–1670. https://doi.org/10.1007/s00477-016-1369-5

Yoon, H., Jun, S.-C., Hyun, Y., Bae, G.-O., Lee, K.-K., 2011. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. J. Hydrol. 396, 128–138.

## 3.2 Groundwater level forecasting using linear time series modeling: the case study of the thermal aquifer system of Monsummano Terme (central Italy).

Chiara Zanotti *[1], Tullia Bonomi [1], Marco S. Nannucci [2], Marco Rotiroti [1]

(1) Department of Earth and Environmental Sciences, University of Milano-Bicocca, Milan, Italy
(2) Regione Toscana, Direzione Ambiente ed Energia, Firenze, Italy

*Corresponding author; Address: Piazza della Scienza 1, 20126 Milano, Italy; Tel: +39 0264482882; Fax: +39 02 64482895; e-mail: c.zanotti@campus.unimib.it

**Rend. Online Soc. Geol. It., Vol. 47 (2019), pp. 153-160**
**Doi:** https://doi.org/10.3301/ROL.2019.27

### 3.2.1 Abstract

Groundwater level forecasting can support a more efficient and sustainable groundwater management. In the present study, linear models are implemented on hydrogeological and meteorological time series related to the thermal aquifer system of Monsummano Terme (Pistoia province, Tuscany region in central Italy), in order to develop a tool for short-term groundwater level forecasting.

Groundwater level data were analyzed along with daily precipitation and used for training Autoregressive (AR) models and Autoregressive models with Exogeous Input (ARx). The best performing AR model consisted in a linear combination of the past 9 groundwater level measurements whereas the ARx model concerned, in addition, the precipitation values of the past 10 days. As a result, these models show an optimal performance in terms of Normalized Root Mean Square, and their residuals (i.e. modeled – measured values) can be associated to a white noise containing no relevant information. This study shows how linear models can be useful and easy-applicable tools for groundwater level forecasting in systems characterized by a linear relationship between recharge and groundwater level increase.

KEY WORDS: Time Series Analysis, AR, ARx, Exploratory Analysis, Residual Diagnostics.

### 3.2.2 Introduction

Groundwater level modeling is a useful tool to support groundwater resource management (Prinos et al., 2002). Beyond sustaining an efficient and sustainable water exploitation, groundwater level modeling and forecasting can provide a better understanding of the dynamics governing groundwater levels oscillations (Moosavi et al., 2013). Conceptual or physically based models are widely considered the main type of model used to represent hydrological variables and their relationships in terms of physical processes occurring in a particular system (Adamowski and Chan, 2011). At the same time, the increasing availability of detailed hydraulic heads temporal data, and the lack of geological and lithological data led to an increasing attention toward data-driven time series analysis methods.

Several modeling approaches were presented in literature. More than a few studies performed innovative non-linear black-box models such as Artificial Neural Network (Coppola et al., 2003; Daliakopoulos et al., 2005; Nayak et al., 2006; Mohanty et al., 2010; Shiri & Kişi, 2011; Izady et al., 2013); however, some authors (e.g. Beriro et al., 2013) argued that the application of a linear model can be more appropriate in simple systems characterized by linear relationships between the hydrological variables.

This work focuses on groundwater level forecasting in an aquifer system where a linear relationship between recharge by rainfall and hydraulic head response can be assumed. This aquifer system consists in a fractured aquifer which can be considered as a closed system. That might be the case where a linear model can successfully capture the dynamics of the system. The purpose of this study is to test the effectiveness of linear models, such as Auto Regressive model (AR) and Autoregressive model with exogenous input (ARx), in forecasting groundwater levels when a linear relation between groundwater response and recharge can be assumed. In conjunction with this main goal two other key aspects were taken into consideration and highlighted, wich are (a) the usefullness of the exploratory analysis as a basis for the model choice and implementation and (b) the importance of combining quantitative and qualitative error analysis.

### 3.2.3 Study area

This study involves the thermal aquifer system of Monsummano Terme in central Italy (Pistoia province, Tuscany Region). This system is located at the transition between the northern Apennines and the Arno Plain and consists of a fractured carbonate aquifer.

A previous study (Grassi et al., 2011) pointed out that the Monsumanno Terme aquifer has a short extension (~1-2 km of length and width and 600-700 m of thickness) and can be considered as a closed system which is indipendent and physically separated from the nearby Montecatini Terme aquifer system and has no interrelations with surface water bodies. The system is recharged upstream by local precipitation in an hilly area where carbonates outcrop (Monsummano Alto, at 150-300 m a.s.l approximately) and discharges downstream through natural springs and well abstractions (Monsummano Terme, at ~50 m a.s.l.). More specifically, the conceptual model for water circulation in the Monsummano Terme system (Grassi et al., 2011) considers that, after infiltrating in the recharge area, water mainly moves downward toward the aquifer basement (at 600-700 m of depth) due to the predominant vertical fracturing. Successively, the water flow assumes a predominant orizontal component toward the discharge area; during this circulation, the water increases its temperature due to the interactions with deep thermal fluids. Finally, in proximity of the discharge area, the water mainly moves upward due to a) its lower density caused by the temperature increase and b) the presence of low-permeability deposits at the transition between carbonates and alluvial deposits.

Grassi et al. (2011) identified a linear and quick response of the groundwater levels in the discharge area to the precipitation in the rechage area.

The features described above allow to consider the Monsummano Terme system as a proper case study in order to test the groundwater level forecasting capability of linear models.

### 3.2.4 Materials and methods

This work was carried in two phases consisting in a) an exploratory analysis and b) the implementation and evaluation of linear models. In each phase different statistical techniques were applied with a specific objective. In the exploratory analysis Auto Correlation, Partial

Autocorrelation, Cross Correlation and Impulse Response were applied, aiming at understanding the information contained within the available data. In the second phase AR and ARx (Box & Jenkins, 1979) models were implemented and evaluated as tools to perform short term forecasting of the groundwater level.

#### *3.2.4.1 Available data*

The present study is based on two available time series: 1) a sequence of daily groundwater levels (Fig. 1, c, black line) measured at the monitoring well of Grotta Giusti (Fig. 1, a), located in the discharge area of the Monsummano Terme system and 2) a sequence of daily precipitation (Fig. 1, c, blue line) registered at the Montecatini Terme station, that is ~5 km far from Monsummano and can be considered representative of the precipitation in the recharge area of the Monsummano Terme system. Both time series cover a period of almost 10 years, from 26/12/2005 to 07/05/2015.

The monitoring well at Grotta Giusti is 158 m deep and filtered from about 70 m of depth. The average groundwater level on the considered period is 57.86 m a.s.l., with a standard deviation of 1.18 m. Since 2008 (Fig. 1, c), monitoring showed a significant monthly variability that was not found in the first period from 2005 to 2008. A seasonal pattern is evident, with the minimum in summer months and maximum (with daily peaks) between November and December.

Precipitation (Fig. 1,c) shows a seasonality as well, but it is stationary on the considered period as there is no trend; the wettest months are in the period between November and January, while in the period between June and August the frequency of rainy days decreases. The rainiest year in the considered series is 2014 with a cumulative precipitation amount of 1662 mm, while the driest is 2011 with 848 mm.

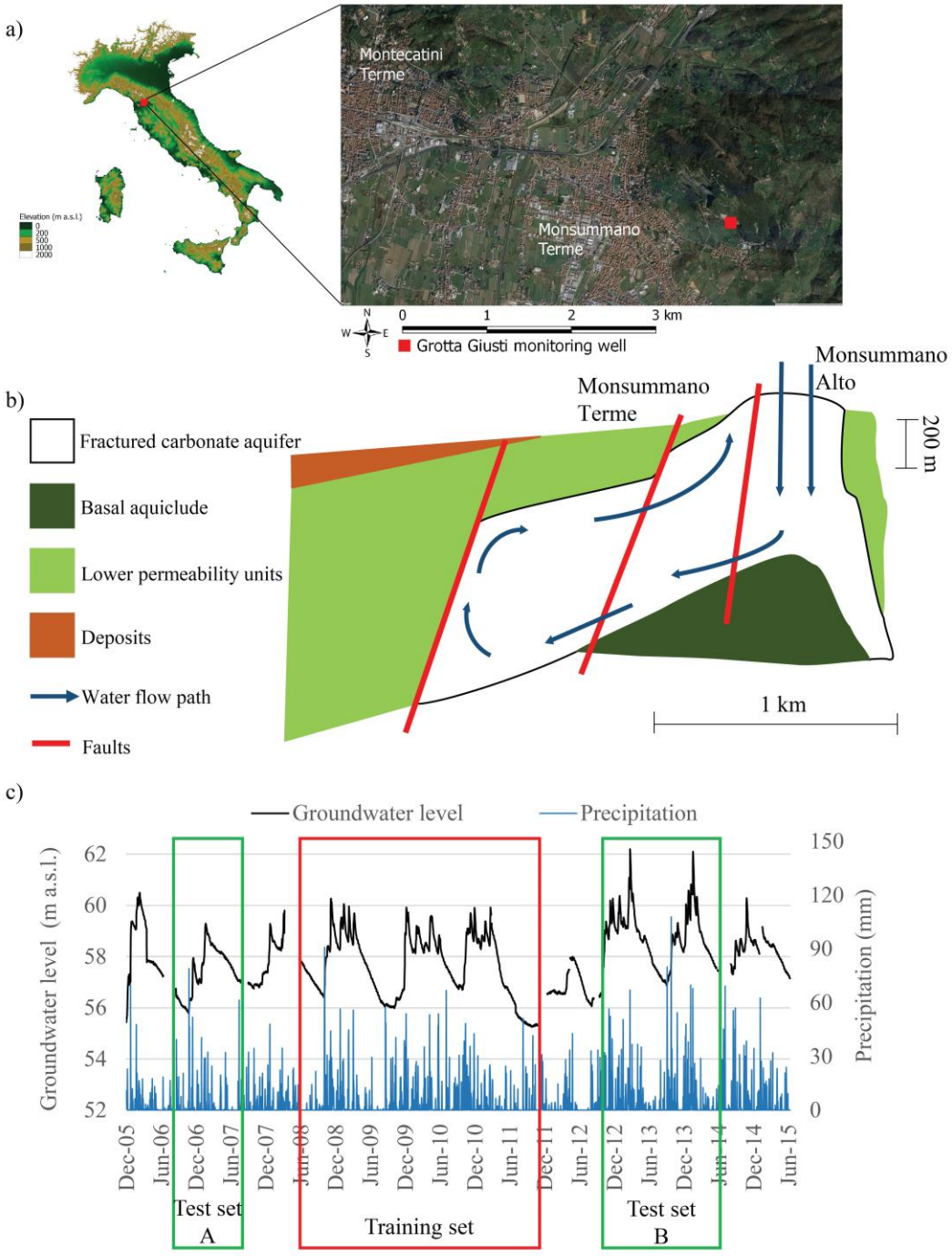



**Figure 1 - a) Location of Grotta Giusti monitoring well. b) Schematic representation of the aquifer and water flow paths according with the conceptual model presented by Grassi et al. 2011. c) Groundwater level and precipitation time series**

Within the groundwater levels sequence, a certain number of missing data were pinpointed. The development of auto regressive models requires continuous sequences of data, without any missing data. Similarly, for the validation of these models, it is appropriate to apply them to sequences of continuous data in order to evaluate their quality and their forecasting power. Since the absence of some data is a common condition in almost all the environmental data sets, techniques have been developed, such as interpolation, able to reconstruct the gaps.

The Fig. 1 (c) shows the missing data within the available groundwater level sequence and how they are grouped within circumscribed time windows, whereas suitable subsequences of continuous data are recognizable. Therefore, the three longest continuous subsequences were identified to process the data (Fig. 1, c). In this way, it was possible to avoid dealing with reconstructed data, basing the analysis only on actually measured data. The three sequences were used at different stages of the analysis and development of the regression models. In particular, the most populated sequence was used as training set while the two remaining sequences, shorter than the first one, were successively used as test sets in order to evaluate the performance of the model trained on the training set:

3) Training set: 1246 groundwater level measurements from 19/06/2008 to 16/11/2011 with an average value and standard deviation of 57.62 m a.s.l. and 1.22 m, respectively; the average daily precipitation is 3.45 mm.

4) Test set A: 354 groundwater level measurements from 12/09/2006 to 30/08/2007 with an average value and standard deviation of 57.40 m a.s.l. and 0.89 m, respectively; the average daily precipitation is 2.84 mm.

5) Test set B: a total number of 613 groundwater level measurements from 12/10/2012 to 16/06/2014 with an average value and standard deviation of 58.71 m a.s.l. and 1.08 m, respectively; the average daily precipitation is 4.55 mm.

#### *3.2.4.2 Time-series exploratory analysis*

The first phase of the present work consisted in exploratory analysis, aimed at identifying the information contained within the available data, in order to steer the models implementation. Since the studied data consist in time-domain sequences, the exploratory analysis was

conducted by means of techniques belonging to the family of the time-series analysis. More specifically, Auto Correlation and Partial Autocorrelation were applied as univariate statistics, whereas Cross Correlation and Impulse Response were used as bivariate statistics.

Time series are characterized by autocorrelated data: given the groundwater level of the *k*-1 day, it can be stated that the groundwater level of the next day *k* falls within a set of values close to the one of the day *k*-1, so a totally random value cannot be assumed. On the other hand, it is also plausible to assume that the groundwater level of the day *k* depends only on a finite number of previous values. The Auto Correlation is a measurement of this relationship between different values of the same variable (in this case groundwater level) at different time: it represents the correlation between the groundwater level detected at a given time *k* and the groundwater level data collected at previous timesteps. Autocorrelation values close to zero indicate that there is no correlation while values close to 1 or -1 (or higher than a significance level) indicate strong positive or negative correlations.

The formal definition of Auto Correlation (Eqn. 1) considers the linear dependence of a feature with itself at different time lags. A process can be defined stationary if the Auto Correlation between any two measurements only depends on the number of time lags between them.

$$C(k) = \frac{1}{n}\sum_{t=1}^{n-k} (x_t - \bar{x}) \cdot (x_{t+k} - \bar{x}), k \geq 0 \quad (1)$$

$$\gamma(k) = \frac{C(k)}{C(0)} \quad (2)$$

Where *C(k)* is the correlogram, *n* is the time-series length, *k* is the time-lag, $x_t$ is the value of the variable *x* at time *t*, and $\bar{x}$ is the average of the *x* variable; $\gamma(k)$ is the autocorrelation function.

The Partial Autocorrelation, denoted $\phi_{h,h}$, is the Auto Correlation between $y_t$ and $y_{t-h}$ after removing any linear dependence on $y_{t-1}$, $y_{t-2}$, ..., $y_{t-h+1}$ in order to avoid correlation resulting from

a mutual linear dependence on the time lags between *t* and *t-h* (Box et al., 2015). Partial Autocorrelation values close to zero indicate no correlation, while values higher than the significance level indicate that a correlation exists. Auto Correlation and Partial Autocorrelation functions are sequences of, respectively, $\rho_h$ or $\varphi_h$ for each $h$=1, 2, … , $n$ and they can be useful tools for choosing the model and the model orders (Mishra & Desai, 2005). An AR model can be a suitable solution for systems whose Auto Correlation Function tails off gradually whereas its Partial Auto Correlation Function cuts off after n lags. More specifically, if the values of the Partial Autocorrelation function are zero at lag $n + 1$ and greater the process can be represented with an AR(n) model based on $n$ previous values.

The Cross Correlation is a measure of the linear correlation between two variables at different time-lags. An estimate of the samples Cross Correlation is:

$$C_{xy}(k) = \frac{1}{n}\sum_{t=1}^{n-k} (x_t - \bar{x}) \cdot (y_{t+k} - \bar{y}), k \geq 0 \quad (3)$$

$$\gamma_{xy}(k) = \frac{C_{xy}(k)}{\sigma_x \sigma_y} \quad (4)$$

Where $C_{xy}(k)$ is the cross - correlogram, $n$ is the time-series length, $k$ is the time-lag, $x_t$ is the value of the output variable $x$ (groundwater level in this case) at time $t$, and $y_{t+k}$ is the value of the input variable (precipitation) at time $t+k$; $\bar{x}$ and $\bar{y}$ are the average values of the $x$ and $y$ variables and $\sigma_x$ and $\sigma_y$ the standard deviations. $\gamma_{xy}(k)$ is the Cross Correlation function.

The information obtained from a Cross Correlation graph is smoothened by the effect of the Auto Correlation that is still present within the data time series, in order to avoid this effect, the Impulse Response function has been examined. This function focuses on the relationship between the two variables avoiding the influence of the Auto Correlation, by filtering the input and the output data with an autoregressive polynomial function before calculating the Cross Correlogram in Eqn. 3 (see Signal Processing Toolbox™ User's Guide, R2016a). Both the Cross Correlation and the Impulse Response functions provide information about the

relationship between groundwater level and precipitation at a certain time lag: values close to zero indicate that there is no relationship while values outside the significance boundaries indicate a significant effect of the precipitation on the groundwater level.

#### *3.2.4.3 Linear models*

A system is definable linear if the output generated by the linear combination of two or more inputs is equal to the linear combination of the outputs generated by the single inputs. It is also definable time - invariant when the output generated by a delayed input is equal to the output generated by the delayed original signal (the system response is not directly dependent from the instant t).

Discrete-time and univariate models were applied in the second phase of the present work to process and reconstruct the time series of groundwater level and precipitation of the Monsummano Terme system. These are AR and ARx linear models, both implemented using MATLAB®. The AR and ARx (Eqn. 5 and 6) are both part of the linear models family, and therefore are usually adopted to model linear systems.

Given $k$ the current instant, the purely AR model of the hydraulic heads sequence $y[\cdot]$ (Eqn. 5) states that each $y[k]$ value is a linear combination of the n past values (where n is the order of the model), with coefficients, assumed to be constant. In the development of an ARx (Eqn. 6) the exogenous input (rainfall) has been also considered. The ARx model defines each $y[k]$ value as a linear combination of $n$ previous values of $y$ and $p$ values of the input variable, where $n$ and $p$ values are called orders of the model and reported with the notation [n p].

$$y_t = \sum_{j=1}^{n} \vartheta_j \, y_{t-j} + \varepsilon_t \tag{5}$$

$$y_t = \sum_{j=1}^{n} \vartheta_j \, y_{t-j} + \sum_{j=1}^{p} \varphi_j \, u_{t-j} + \varepsilon_t \tag{6}$$

Where $y_t$ and $u_t$ are, respectively groundwater level and precipitation value at time $t$; θ and φ are the coefficients of the models and ε is a white noise element with 0 mean and variance $\sigma^2$. The parameters n and p are the model orders. For the determination of the coefficients (θ and φ) a non-iterative algorithm has been used, aiming at minimizing least squares errors.

The definition of the orders of the models, respectively, n and [n p] is based on "*a priori*" considerations on the data and on the conceptual model and "*a posteriori*" considerations based on comparing the effectiveness of AR and ARx models with different orders.

The AR and ARx models where trained on the training set and their effectiveness was the determined by applying them on the test sets. The quality of the elaborated AR and ARx models was evaluated in terms of Normalized Root-Mean-Square Error (NRMSE) (Eqn. 7):

$$NRMSE = 100\left(1 - \sqrt[2]{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - mean(y))}}\right) \quad (7)$$

A second criterion has been taken into consideration as a tool to compare different models which is the normalized Akaike Information Criterion (nAIC); nAIC unifies in one value the information about the effectiveness of the model and its complexity, penalizing either bad accuracy and excessive complexity, as expressed by Eqn. 8:

$$nAIC = \log\left(det\left(\frac{1}{n}E^T E\right)\right) + \frac{2 * n_p}{n} \quad (8)$$

Where $E$ is the $N$-by-$1$ matrix of prediction errors, $n_p$ is the number of free parameters of the model and $n$ is the number of data in the dataset. The best model has the lowest nAIC value.

Besides the effective performance of the models expressed as NRMSE and AIC, the choice of the model order was based on a detailed analysis of the model errors, aimed at ensuring that no relevant information was contained in the residual of the model (Ginocchi et al., 2016). Specifically, all the techniques used in the exploratory analysis of the data sequences (Auto Correlation Function, Partial Autocorrelation Function, Cross Correlation and Impulse Response) were applied on the residual of the models. Residuals were calculated for each sample as the difference between the measured and modelled values. A model can be

considered effective when the error is associable to a white noise (i.e. its residuals have no significant Auto Correlation or Cross Correlation with its input).

### 3.2.5 Results and discussion

#### *3.2.5.1 Exploratory analysis*

Exploratory analysis was performed in the first phase of the present work, in order to determine how much the groundwater level at a certain time *t* is affected by previous values of groundwater level (Auto Correlation and Partial Autocorrelation) and past precipitation events (Cross Correlation and Impulse Response).

Fig. 2 (a,b) shows the Auto Correlation and Partial Autocorrelation Functions of the groundwater level data and the Cross Correlation and Impulse Response between groundwater level and precipitation. For each graph the estimate of the function for 15 time lags (days) is represented, and the significance level with a 95% confidence is shown as a horizontal line. Fig. 2 (a) shows that a strong Auto Correlation (outside the significance boundaries) exists within all the first 15 lags, slowly decreasing from $t_0$ to $t_{-15}$, while the Partial Autocorrelation reaches significant values until the 9$^{th}$ lag.

As regards the input-output relationship between precipitation and groundwater level, the Cross Correlation and Impulse Response (Fig. 2 c and d) resulted with similar trends. The Cross Correlation is smoothened by the implicit effect of the data auto-correlation which means that the groundwater level at a certain time is affected by the precipitation but also by previous days groundwater levels. The Impulse Response function, that is a cross correlation calculated after filtering out the effect the autocorrelation of the data, shows a more evident cut off after the 10$^{th}$ lag when it decreases below the significance limit.

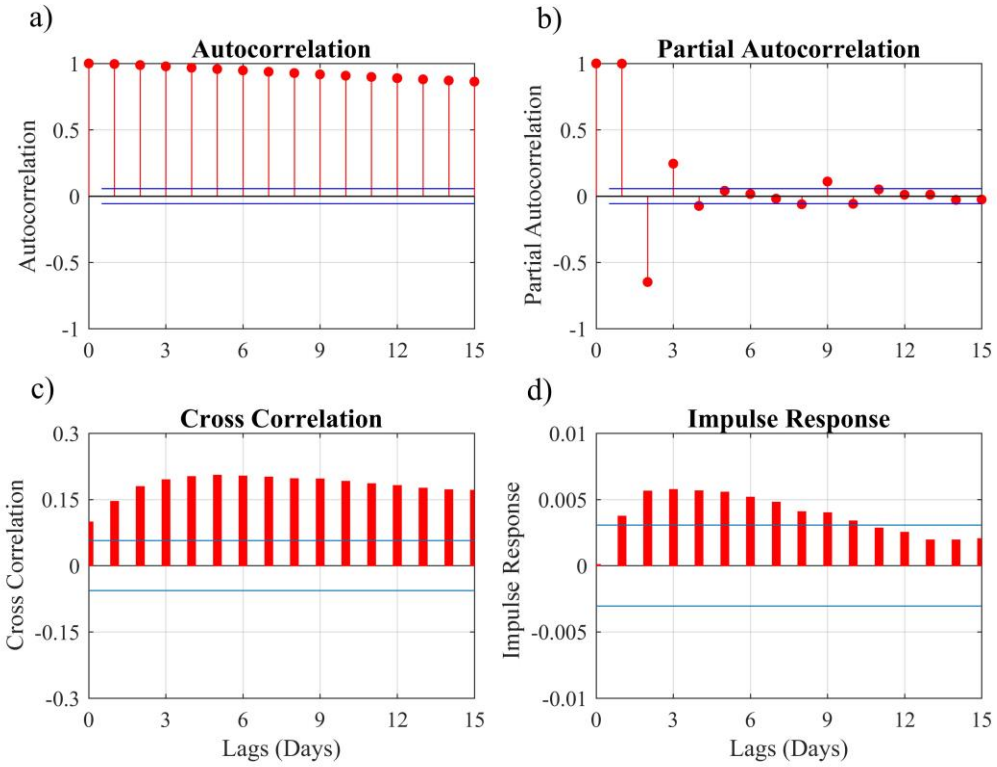



**Figure 2 a) Auto Correlation function b) Partial Autocorrelation Function c) Cross Correlation Function d) Impulse Response Function**

The resulted trends of, respectively, Auto Correlation and Partial Autocorrelation, Cross Correlation and Impulse Response support the original hypothesis of the system linearity and suggest that a AR and ARx models can be appropriately applied to the available data. Particularly, the results of Partial Autocorrelation suggest that the groundwater level measured at a certain time is affected by 9 past values of groundwater level while the Impulse Response indicates that it is also affected by 10 past precipitation values, leading to hypothesize that an AR(9) model and an ARx([9 10]) model could appropriately represent the system.

#### *3.2.5.2 Implementation of linear models*

In this phase AR models with different orders (up to 15) were elaborated on the training set and compared by applying them on both training and test sets, in order to determine whether

the information obtained by the exploratory analysis could actually lead to the best performing model.

Tab. 1 shows the fitting power, in terms of NRMSE, of the AR models with increasing orders elaborated on the training set. Fitting values are reported for both the training and the test sets.

**Table 1 – Effectiveness of AR(1)-AR(15) models in terms of NRMSE**

| Model order | Fitting on Training Set | Fitting on Test Set A | Fitting on Test Set B |
|---|---|---|---|
| **1** | 92.36 | 92.16 | 87.27 |
| **2** | 94.18 | 94 | 89.02 |
| **3** | 94.35 | 94.06 | 89.22 |
| **4** | 94.37 | 94.08 | 89.25 |
| **5** | 94.37 | 94.08 | **89.25** |
| **6** | 94.37 | 94.08 | 89.24 |
| **7** | 94.38 | 94.08 | 89.23 |
| **8** | 94.39 | 94.1 | 89.2 |
| **9** | 94.43 | 94.17 | 89.17 |
| **10** | 94.44 | 94.21 | 89.16 |
| **11** | 94.44 | 94.23 | 89.12 |
| **12** | 94.44 | 94.23 | 89.11 |
| **13** | 94.44 | 94.23 | 89.13 |
| **14** | 94.44 | 94.24 | 89.13 |
| **15** | **94.45** | **94.25** | 89.13 |

Results show that the fitting on the training set increases progressively from 92.36% to 94.45%, from a model of order 1 to a model of order 15. The forecasting fitting power on the test sets instead grows uninterrupted in the case of the test set A (implying that the reversal point concerns even higher orders) whereas the test set B shows a maximum for the model of order 5.

In general, it is expected that the value of the fitting on the reconstruction of the training set gradually increases with increasing order of the models, whereas the fitting on the prediction of different values (test set) increases in a first phase but it drops once an optimum has been reached. On this basis, our results suggest that the AR(5) model would offer the best predictive power. However, the residuals analysis of AR(5) shows values of Auto Correlation higher than the significance level (horizontal line) for lags 6, 7 and 14 (Fig. 3, a), so that these residuals cannot be considered as a white noise and that means that the model is not representing correctly the Auto Correlation of the data, so that it cannot be considered as the best performing model. Residual Analysis of the models with higher orders, on the other hand, indicate that the AR(9) model is the one that couples higher NRMSE (Tab. 1) values with not significant residuals Auto Correlation (Fig. 3, b).

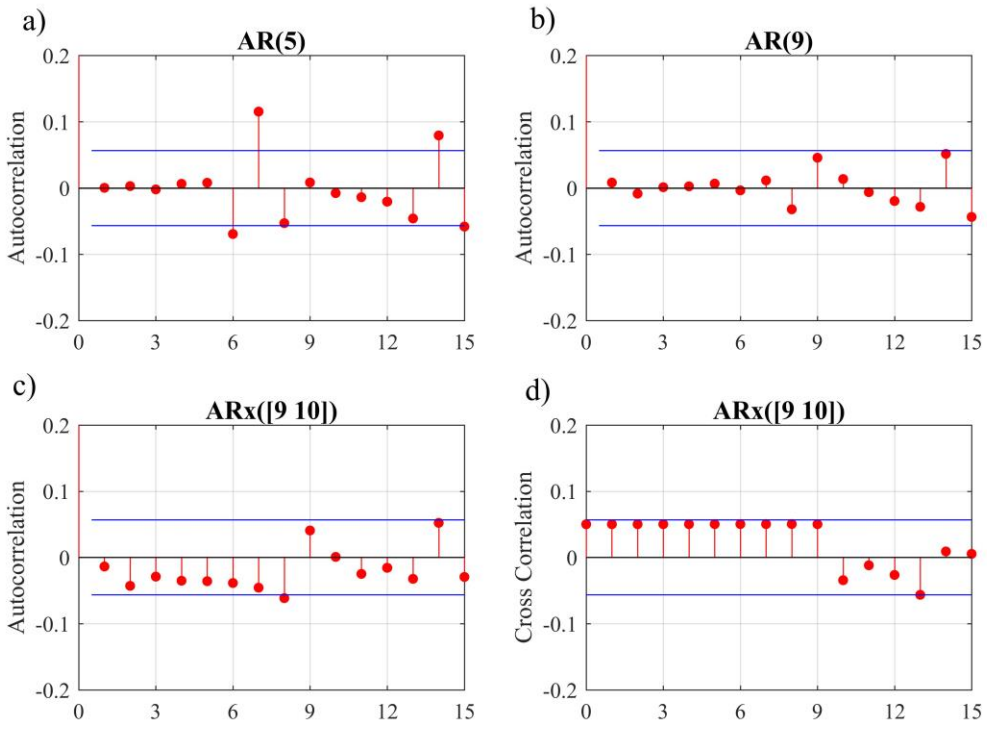



**Figure 3 - a) Residual autocorrelation of AR(5). b) Residual autocorrelation of AR(9). c) Residual autocorrelation of ARx([9 10]). d) Residual and precipitation Cross Correlation of ARx([9 10]).**

As for the ARx model, the order of the Auto Regressive order in the ARx model has been maintained as stated for the AR model, while the choice of the order of the exogenous part (i.e. the precipitation input) was based on the trend and the significance level of the Impulse Response Function that stated that the groundwater level of a certain day is affected by 10 past precipitation days.

The evaluation of the NRMSE confirms that the ARx([9 10]) model, that is a linear combination of 9 past groundwater level and 10 precipitation values, can successfully capture and represent the correlation between precipitation and groundwater level, reaching a fitting value of 95.22% on the one step ahead prediction on the training set, while 94.4% and 90.67% on respectively test set A and B.

Fig. 3 (c,d) shows the Auto Correlation of the residuals of the ARx([9 10]) model on the training set, and the Cross Correlation between residuals and precipitation: both functions are not significant for each considered lag, indicating that the model error is associable to a white noise, confirming the good performance of the ARx([9 10]) model.

As regards the nAIC criterion, AR(5), AR(9) and ARx([9 10]) have a nAIC value of respectively: -5.33, -5.34 and -5.62. It confirms that generally the AR(9) outperforms AR(5) and that the ARx model outperforms the AR and that the increase of parameters is justified by an increase of the accuracy of the model.

Both AR and ARx models have been successively tested on wider forecasting horizons, in order to evaluate whether they could be considered reliable also for longer predictions. The predictive power decreases as the forecasting horizons increases. That means that these models can be very accurate on the short-term predictions (1-3 days) while their reliability decreases on the long-term predictions.

In terms of the three step ahead forecasting for example, the AR(9) model NRMSE are 83.7%, 83.28% and 71.41% for, respectively, training set, test set A and test set B.

Also in the case of the three step ahead prediction, the ARx model outperforms the AR model reaching NRMSE values of 87.14%, 85.35% and 78.51% for respectively training set, test set A and test set B.

### 3.2.6 Conclusions

In this work, linear models were applied to analyze groundwater level and precipitation time series. Results showed that these models could be used as valid and easily applicable tools to perform short term groundwater level forecasting in systems characterized by a linear relationship between recharge and water table rising, as the case of the Monsummano Terme system, Pistoia province, in Tuscany region. In general, predictions obtained with data driven techniques are more accurate within observed limits; in the present study the models have been applied on a test set where the groundwater level exceeds the range of the training set, leading to a more accurate quantification of the model accuracy outside the range of its training domain.

In the case study the AR(9) model shows the best predictive power among AR models with different orders, whereas within the ARx models the ARx([9 10]) outperforms the others. Both AR(9) and ARx([9 10]) showed high performances on both one or three days ahead predictions where in both cases the one day ahead prediction resulted more accurate than the three days ahead one. The performance of the models on the test set B that has values outside the range of the training set slightly decreases compared to the accuracy on the training set but the accuracy is still high.

Furthermore, this study highlights the importance of exploratory analysis that can constitute a solid and meaningful basis for choosing model type and model orders. In particular, the exploratory analysis leads to the following remarks:

- data characterized by a smoothly decreasing Auto Correlation Function, whose Partial Autocorrelation Function cuts off after a certain value can be represented by AR;
- Partial Autocorrelation Function offers an effective indication of the AR model order;
- Cross Correlation Function and Impulse Response can support the implementation of an ARx model by graphically highlighting the range in which can follow the relevant number of past precipitation level.

Finally, this study underlines the importance of a detailed residual analysis: error functions such as NRMSE or AIC do not provide alone an exhaustive indication of the model performance since they only consider residuals quantitatively. On the other hand, Auto Correlation Function of the residuals and the Cross Correlation between residuals and the input of the model (in this case precipitation) can actually help understanding whether the error still contains relevant information, a condition that should be avoided.

### 3.2.7 References

Adamowski J., Chan H.F. (2011) - A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol., 407, 28–40.

Beriro D.J., Abrahart R.J., Paul Nathanail C. (2013) – Letter to the editor on "Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations" by Jalal Shiri & Ozgur Kisi [Computers and Geosciences (2011) 1692-1701]. Comput. Geosci., 56, 216–220.

Box G.E.P & Jenkins G.M. (1979) - Time Series Analysis: Forecasting and Control. Holden-Day, Oakland, CA.

Coppola E., Poulton M., Charles E., Dustman J., Szidarovszky F. (2003) - Application of Artificial Neural Networks to Complex Groundwater Management Problems. Nat. Resour. Res., 12, 320.

Daliakopoulos I.N., Coulibaly P., Tsanis I.K. (2005) - Groundwater level forecasting using artificial neural networks. J. Hydrol., 309, 229–240.

Ginocchi M., Franco G.F.., Rotiroti M., Bonomi T. (2016) - Analysis and prediction of groundwater level time series with Autoregressive Linear Models. Rend. Online Soc. Geol. It., 39, 109–112.

Grassi S., Doveri P.M., Ellero A., Palmieri F., Vaselli L., Calvi E., Trifirò S. (2011) - Studio dei sistemi termali di Montecatini e Monsummano Terme. Consiglio Nazionale delle Ricerche – Istituto di Geoscienze e GeorisorseTechnical Report, 109 pp.

Izady A., Davary K., Alizadeh A., Moghaddam Nia A., Ziaei A.N., Hasheminia S.M. (2013) - Application of NN-ARX Model to Predict Groundwater Levels in the Neishaboor Plain, Iran. Water Resour. Manag., 27, 4773–4794.

Mishra A.K., Desai V.R. (2005) - Drought forecasting using stochastic models. Stoch. Environ. Res. Risk Assess., 19, 326–339.

Mohanty S., Jha M.K., Kumar A., Sudheer K.P., (2010) - Artificial neural network modeling for groundwater level forecasting in a river island of eastern India. Water Resour. Manag., 24, 1845–1865.

Moosavi V., Vafakhah M., Shirmohammadi B., Behnia N. (2013) - A Wavelet-ANFIS Hybrid Model for Groundwater Level Forecasting for Different Prediction Periods. Water Resour. Manag., 27, 1301–1321.

Nayak P.C., Satyaji Rao Y.R., Sudheer K.P. (2006) - Groundwater level forecasting in a shallow aquifer using artificial neural network approach. Water Resour. Manag., 20, 77–90.

Prinos S.T., Lietz A.C., Irvin R.B. (2002) - Design of a real-time ground-water level monitoring network and portrayal of hydrologic data in southern Florida. U.S. Geological Survey, Water-Resources Investigations Report 2001-4275, 115 pp.

Shiri J., Kişi Ö. (2011) - Comparison of genetic programming with neuro-fuzzy systems for predicting short-term water table depth fluctuations. Comput. Geosci. 37, 1692–1701.

The MathWorks, I., 2016. Signal Processing Toolbox TM User's Guide.

## 3.3 Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system

Chiara Zanotti[1]*, Marco Rotiroti[1], Simone Sterlacchini[2], Giacomo Cappellini[2], Letizia Fumagalli[1], Gennaro A. Stefania[1], Marco S. Nannucci[3], Barbara Leoni[1], Tullia Bonomi[1]

[1] Department of Earth and Environmental Sciences, University of Milano-Bicocca, Piazza della Scienza, 1, 20126 Milano, Italy. Email: c.zanotti@campus.unimib.it

[2] National Research Council, Institute for the Dynamics of Environmental Processes, Piazza della Scienza, 1, 20126 Milano, Italy

[3] Regione Toscana, Direzione Ambiente ed Energia, Settore Servizi Pubblici Locali, Via San Gallo 34/a, 50129 Firenze, Italy

**Journal of Hydrology Vol. 578 (2019), n. 124015**

**Doi:** https://doi.org/10.1016/j.jhydrol.2019.124015

### 3.3.1 Abstract

Groundwater level forecasting is a useful tool for a more efficient and sustainable groundwater resource management. Developing models that can accurately reproduce groundwater level response to meteorological conditions can lead to a better understanding of the groundwater resource availability. Here an autoregressive neural network (NNARx) approach is proposed and compared with autoregressive linear models with exogenous input (ARx) in order to forecast groundwater level in an aquifer system where a linear groundwater level response to recharge by rainfall is observed. A well known problem regarding neural networks consists in the high risk of overfitting. Here, three NNARx model were trained using different methods to avoid overfitting: Early stopping, Bayesian regularization and a combination of both. The results show that on the short term forecasting (up to 15 days) the performance of NNARx and ARx are comparable but the ARx model generalizes better, while the NNARx trained with Bayesian regularization outperforms the linear models and the other NNARx models on longer scenarios on the test set. As linear models are less time demanding and do not require

high computational power, they can be considered as suitable tools for short term groundwater level forecasting in linear systems while when longer scenarios are needed neural networks can be considered more reliable, and training them with Bayesian regularization allows to minimize the risk of overfitting.

KEYWORDS: Bayesian Regularization; Groundwater level forecasting; Linear model; Neural Networks; Overfitting.

### 3.3.2 1. Introduction

Groundwater level forecasting is a useful tool for supporting the sustainable management of water resources. Particularly, short term forecasting (few days ahead) can help the assessment of groundwater resources availability for fulfilling human water needs while long term forecasting can be useful in terms of running scenarios for a changing climate.

In the last decades the attention kept growing toward data-driven techniques as tools to perform groundwater level forecasting, as alternatives to physically-based models. A main advantage of data-driven models with respect to physically-based models is that they do not require information on the hydrogeological parameters of the aquifer (Adamowski and Chan, 2011), that also need to be accurately calibrated (Burrows and Doherty, 2015, 2016; Gianni et al., 2019; Stefania et al., 2018). Therefore, data-driven models can be more suitable when exhaustive data on subsurface properties are not available. Moreover, data-driven models have been proven to be preferable to physically-based models for simulating and forecasting groundwater level variations, since they can overcome data limitations, parameter uncertainty and the challenging problems related to the applicability of physically-based models (Banerjee et al., 2009; Juan et al., 2015; Maskey et al., 2000; Nikolos et al., 2008; Yoon et al., 2011; Zhang et al., 2018). The successful application of data-driven models for forecasting groundwater level variations is shown by many previous works that demonstrated their applicability in a variety of hydrological settings and hydrogeological contexts (Wunsch et al., 2018), such as hard rock systems (Banerjee et al., 2009), river islands (Mohanty et al., 2010), costal aquifers (Taormina et al., 2012; Yoon et al., 2011) and arid zones (Yang et al., 2009).

Data-driven models have, however, some limitations with respect to physically-based models. A main disadvantage of data driven model is that they require long time-series of hydrological and meteorological data to be trained. Other limitations are related to the quantification of the uncertainty and the management of overfitting (discussed in detail below).

Among data-driven models, two widely used techniques for groundwater level forecasting are a) neural networks and b) linear models. Neural networks consist in black-box models which can provide relatively accurate predictions of groundwater levels (Coppola et al., 2013; Huang et al., 2013; Izady et al., 2013; Khalil et al., 2015; Nayak et al., 2006; Shirmohammadi et al., 2013; Tsanis et al., 2008; Uddameri, 2007). Advantages of neural networks are that they can better handle dynamic behaviors and non-linearity (Adamowski and Chan, 2011; Daliakopoulos et al., 2005).

A well-known problem regarding the development of neural networks consists in avoiding overfitting, which means improving the network generalization. Overfitting is defined as the problem that occurs when a network has been trained too hard to fit the training data, resulting in a network which learned to reproduce noise and peculiarities in the training data rather than to find a general predictive rule (Dietterich, 1995). If overfitting occurs the accuracy of the developed models may deteriorate significantly outside the range of the recharge and hydrogeological conditions that prevailed during the training period (Khalil et al., 2015). Generalization is the opposite of overfitting: a network generalizes well when the accuracy of the network output is acceptable also when working with data that are not included in the training set (Rosin and Fierens, 1995). To have a measure of the overfitting problem and to compare different models, the performance of the model is usually measured on data which were not part of the training set (Kang et al., 2018; Srivastava et al., 2014; Wan et al., 2013; Zaremba et al., 2014).

Currently, avoiding overfitting and improving the generalization of neural networks is a crucial challenge in the development of groundwater level forecasting models, since an increase of hydroclimatic extremes can be expected in mid-latitude regions, such as the Mediterranean (Hartmann et al., 2014). The occurrence of hydroclimatic extremes leads to hydroclimatic conditions which exceed the range of the past data that were used to develop forecasting models. Therefore, to test the reliability of data-driven models for extrapolative prediction it

becomes important to test the model on observational data or time period representing more extreme conditions than those in the training set (Bennett et al., 2013; Harmel et al., 2014).

Moreover, other disadvantages of neural networks are that they require several parameters to be chosen and tested carefully in order to produce good results and testing a wide variety of different parameters could become computationally expensive. Examples embrace network type, size and architecture, neuron activation functions, learning algorithms, stop criterion, loss functions, etc. Furthermore, neural networks are often trained starting from random weights and biases, this means that there is no guarantee of converging to the global minima of the error function leading to a higher uncertainty and instability of the models (Sudheer and Kasiviswanathan, 2017).

Linear models as well, have been widely used to model groundwater level variations in several hydrological contexts (Ahn and Salas, 1997; Ginocchi et al., 2016; Lee et al., 2009; Valipour et al., 2013). Advantages of linear models are that their training demands less time and computational effort. However, the main disadvantage is that they can struggle handling non-linearity. Another limitation of linear models is that they can be properly applied only under the following assumptions: a) stationarity of the input time series and b) the residuals must be a white noise (i.e. their mean should be zero and there should be no autocorrelation and no cross correlation between the input and the residuals) (Choubin and Malekian, 2017). Thus, several authors tend to neglect linear models because they may not always perform well when applied on hydrological time series which are often nonlinear (Khalil et al., 2015; Shirmohammadi et al., 2013; Tokar and Johnson, 1999). However, other authors (Beriro et al., 2012; Choubin and Malekian, 2017) argued that linear models can be more appropriate to model simple systems characterized by linear relationships between the hydrological variables.

This work presents the modelling of groundwater level fluctuations in a fractured carbonate aquifer hosting thermal waters in Tuscany region, central Italy, where a linear hydraulic head response to recharge by rainfall was previously observed (Grassi et al., 2011). Thus, this might be the case where a linear model can successfully capture the dynamics of the system.

The main aims of this work are a) to compare the effectiveness of a linear Autoregressive model with exogenous input (ARx) with Neural Network – ARx combined models (NNARx)

on short and long-term forecasting and b) to test different methods to avoid overfitting for the NNARx models. To address the latter, the Levenberg-Marquardt algorithm, which has been proven to be a fast and reliable method to train neural networks for time series forecasting (Adamowski and Chan, 2011; Adamowski and Karapataki, 2010; Daliakopoulos et al., 2005; Sreekanth et al., 2009), is coupled with two different strategies to avoid overfitting, that are the early stopping and Bayesian regularization. Results are compared in order to understand which strategy achieves a better performance on data outside the range of the training set compared with the other strategies and with the linear models ARx.

Avoiding overfitting (and thus improving model generalization) is a key challenge for modelling groundwater fluctuations in the study area since here hydroclimatic extremes could cause problems for the local inhabitants. Indeed, flooding of underground structures due to an inefficient drainage of groundwater has been reported in the study area and this can be aggravated by intense precipitations leading to groundwater level rises. On the other hand, the opposite extreme (i.e. drought periods) together with an increase in groundwater demand for thermal baths could lead to subsidence phenomena with damage to buildings, as happened in other sites in Tuscany region. Local authorities urge to analyse the existing groundwater level data to develop tools capable to deal with hydroclimatic extremes for improving their water resources management.

## 3.3.3 Materials and methods

### *3.3.3.1 Study area*

The study covers an area of ~9 $km^2$ located north-east of the village of Monsummano Terme in Tuscany region, central Italy (Fig. 1), at the transition between the northern Apennines and the Arno Plain. The aquifer system consists of a fractured carbonate hosting thermal groundwater.

Owing to its small extension (about 1-2 km long and 600-700 m thick) and the absence of interconnections with surface water bodies and the downstream alluvial aquifer hosted by the Arno Plain (Grassi et al., 2011), the Monsummano Terme aquifer can be considered as a closed

system. This means that the aquifer has a simple groundwater recharge/discharge: the recharge of the system comes mainly by local precipitation in a neighbouring hilly area with carbonate outcrops (Monsummano Alto, at 150-300 m a.s.l approximately) whereas the discharge is downstream through natural springs and well abstractions located nearby the village of Monsummano Terme (20 m a.s.l. approximately).

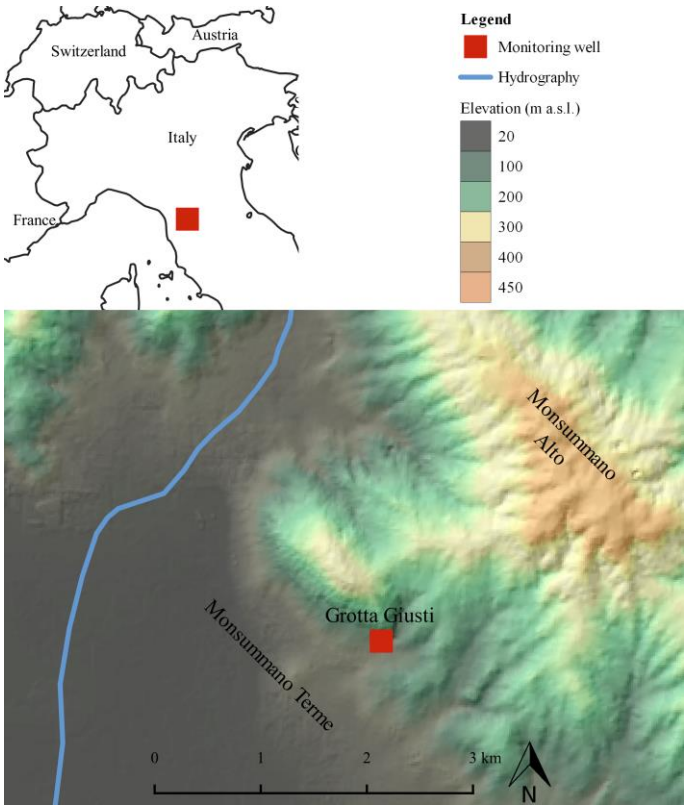



**Figure 1. Study area and location of Grotta Giusti monitoring well.**

In detail, groundwater circulation in the Monsummano Terme system can be described by the following three steps (Grassi et al., 2011). Firstly, precipitation infiltrates in the recharge area with a downward flow direction, toward the aquifer bottom (at 600-700 m depth) due to the predominant vertical fracturing. Successively, the groundwater flow direction assumes a predominant horizontal component, flowing parallel to the aquifer bottom towards the discharge area. In this phase, the interactions with deep thermal fluid determines a

groundwater temperature increase. Finally, in proximity of the discharge area the water reaches lower permeability deposits at the transition between carbonates and alluvial deposits. Consequentially, due to its lower density and the consequential increased temperature, moves upward.

The previous work by Grassi et al. (2011) identified a linear and quick response of the groundwater levels in the discharge area to the precipitation in the recharge area as a result of the immediate pressure propagation within the fractures of the aquifer. This, together with the absence of factors determining nonlinearity in the groundwater response to precipitation such as snow accumulation, strong evapotranspiration and interactions with surface water bodies, allows to consider the Monsummano Terme aquifer as a linear system.

#### *3.3.3.2 Available data*

The data considered in the present study were provided by Regione Toscana that is in charge of protecting water resources. The dataset (Fig. 2) consists in: a) daily groundwater level, automatically measured by a pressure transducer connected to a data-logger at the monitoring well of Grotta Giusti (Fig. 1); b) daily cumulative precipitation and c) maximum temperature registered at the Montecatini Terme station, that is ~5 km far from Monsummano Terme and can be considered as representative of the precipitation and temperature in the recharge area of the Monsummano Terme system.

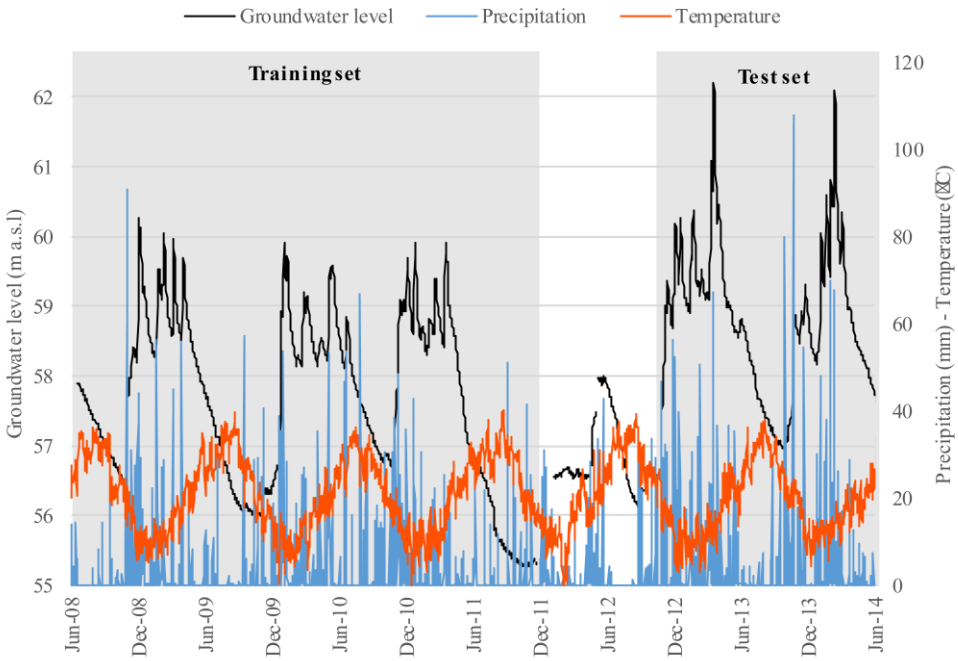



**Figure 2. Available time series of Groundwater level, Precipitation and Temperature.**

#### *3.3.3.3 Forecasting models*

In this study an ARx model has been compared with a NNARx for groundwater level forecasting. Moreover, three NNARx models have been developed with the Levenberg-Marquardt training algorithm and different methods to avoid overfitting during the implementation of the neural networks have been compared. The early stopping method, Bayesian regularization and the combination of both in the training phase of NNARx have been tested. To measure the relative efficiency of the techniques to avoid overfitting, after the training, the models have been applied to a test set (Fig. 2) which contains also extremes outside of the range of the training set. All the models were developed using the neural network time series tool in the MATLAB R2017 software package.

#### *3.3.3.4 Models' inputs*

For the development of the model, one output variable (i.e. groundwater level data) and three input variables were considered: two meteorological variables (i.e. precipitation and temperature) and a time variable.

Several studies consider a time-variable (e.g. time of the day or day of the year) as an input (Candanedo et al., 2018; Lu et al., 2015; Ozbalta et al., 2012; Thomas and Soleimani-Mohseni, 2007). In the present study the information regarding the month of the year was considered by inserting as an input the average groundwater level of all the available data collected in that specific month. That leads to a time-variable with a cyclic trend that has only twelve possible values (i.e. the average groundwater level of each month) repeating over time depending on the month of the year. In this way it has been avoided working with the 1 to 12 month index which would not have a cyclic trend (the months December and January would have index 1 and 12 respectively which have the maximum difference, even if they are temporally close).

Fig. 2 shows the training and test sets, and the trends of the output and input data, allowing to notice the seasonal trends in the groundwater level series. Groundwater levels show a maximum during winter that is the rainiest season, and minimum during summer which is the driest. Several missing data in the groundwater level time series were found between December 2011 and October 2012 due to a malfunction of the measuring instrument. This part of the dataset was excluded from the modelling since continuous series of data are needed to train and test the networks.

Both linear and nonlinear models require choosing a number of previous values of groundwater level and meteorological exogenous inputs (i.e. precipitation and temperature) that are needed for the prediction of the successive time-step. In this study, this choice was based on an in-depth exploratory analysis (Zanotti et al., 2019) which identified the significant number of previous groundwater levels and precipitation values based on the autocorrelation, partial autocorrelation, cross correlation and impulse response. These are techniques aimed at quantifying the "memory effect" that corresponds to the influence of previous value of a variable on successive values of the same variable or of a different one (Chiaudani et al., 2017).

Particularly, the autocorrelation (Fig. S1) function highlights a significant linear correlation between groundwater level values registered at different days. The partial autocorrelation shows that the last significant partial autocorrelation value is at day 9. Therefore, the number of lags (days) considered as model input was 9. Precipitation shows a significant linear cross correlation with the groundwater level. The impulse response, which indicates the cross-correlation after removing the effect of the autocorrelation, indicates that only the previous 10 precipitation days have a significant effect on the groundwater level of a certain day. Since the considered system is linear, the input selection based on linear correlation has been considered valid also for the neural network models even if they would not specifically require linear correlation.

The main physical role of atmospheric temperature and the time variable consists in affecting the relationship between precipitation and groundwater level through the evapotranspiration process; therefore, the number of temperature values considered as model input was kept equal to the precipitation values (i.e. 10). Consequentially, in this study both linear models and neural networks are trained to forecast groundwater level at a certain time based on the previous 9 groundwater levels and 10 precipitation, temperature and time variable values.

#### *3.3.3.5 ARx*

Given $k$ the current instant, the ARx model of the hydraulic heads sequence y[·] (Eqn. 1) states that each groundwater level value at time t, $y_t$ is a linear combination of $n$ previous groundwater level values $y$ and $p$ previous values of the exogenous input variables $u$, with coefficients, $\theta_j$ and $\varphi_j$, assumed to be constant.

$$y_t = \sum_{j=1}^{n} \vartheta_j \, y_{t-j} + \sum_{j=1}^{p} \varphi_j \, u_{t-j} + \varepsilon_t \qquad (1)$$

The residual term $\varepsilon_t$ is a white noise with zero mean. To apply linear models the input time series must be stationary. In this study Phillips–Perron (PP) test was used to assess the stationarity of groundwater level, precipitation, temperature and time variable, and in each case the time series could be considered stationary with a significance level of 0.05, meaning that they could be used without transformations. To prove that the error term $\varepsilon_t$ consists in a white

noise signal, a residual analysis of the ARx model has been performed investigating the average of the residuals, their randomness, their autocorrelation and their cross correlation with the exogenous input variables.

#### *3.3.3.6 NNARx*

Neural networks are flexible computing techniques designed after the biological neuron system. These models have been widely applied in a variety of scientific, technological fields, involving time series analysis, classification, pattern recognition, image processing etc. Several different model structures have been developed in the last decades. The most common neural network structure, called Multilayer Perceptron network (MLP) (Rumelhart and McClelland, 1986) was employed in this study; it consists in one input layer, one output layer and at least one hidden layer (Fig. 3). The defining equation of NNARx model is:

$$y_t = f\big(y(t-1), y(t-2), \dots, y(t-n), y(t-1), i(t-2), \dots, i(t-p)\big) \qquad (2)$$

In this kind of network each processing unit, called neuron (or perceptron), is connected to those in the preceding layers and each connection is characterized by a weight.Inputs *i* coming into the neuron are weighted and summed up, together with a bias element (with its own weight) and only then, they are passed through the activation function of the neuron (Fig. 3).

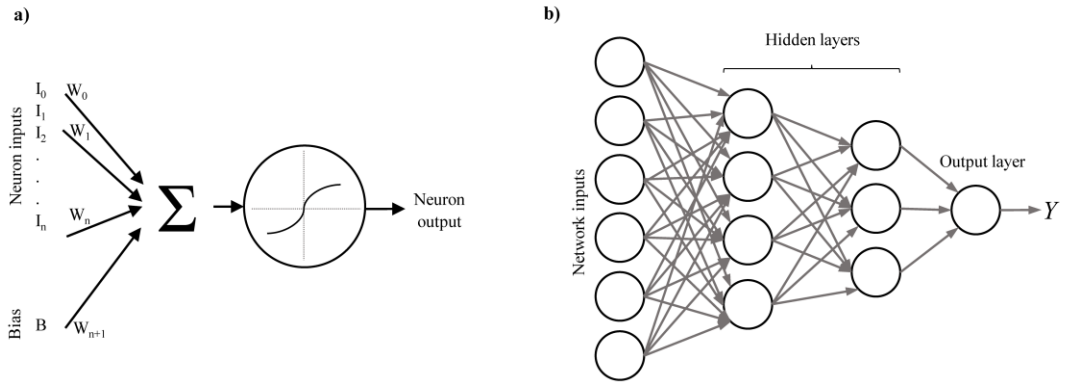



**Figure 3. Schematic representation of a) a neuron, b) a Multilayer Perceptron network (MLP)**

In this study, sigmoid function was used with the exception of the output layer that has linear transfer functions. Particularly, among different kinds of sigmoid activation functions, tan-sigmoid has been proven be the best pertinent transfer function to fit hydrological variables with neural network trained with Levenberg-Marquardt algorithm (Yonaba et al., 2010). In this study a comparison has been performed with linear activation functions and the tan-sigmoid showed the best performance. Results of the comparison are shown in Fig. S2.

To train the networks the Sum of Squared Error (SSE) is minimized:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

Where $y_i$ is the i-th target value of groundwater level, $\hat{y}_i$ is the i-th groundwater level value calculated by the model; n is the total amount of data in the sequence. In feed-forward neural networks data are fed from the input layer to the output layer, through the hidden layers and no flow of information occurs in the opposite direction. In recurrent neural networks, the output of the network can flow from the output layer to the input one, and it can be used as an input to perform multi-step ahead predictions.

To favour the convergence of the network, data were scaled between -1 and 1 which is the range of the chosen activation functions (LeCun et al., 2012).

In this study, after the network has been trained with the feed-forward structure, a connection between the output and the input layer was created leading to a recurrent neural network structure (i.e. closed loop structure). With this re-arranged structure, it is possible to perform multi-step ahead forecasting, by providing to the network only the external input data (i.e. precipitation, temperature and time variable), while the output data (i.e. groundwater level) calculated by the network are inserted again into the network through the input layer to perform the successive predictions. To perform the x step ahead forecasting the measured groundwater level until time t are fed into the network, then the forecast is performed with the closed loop structure until time t+x. This kind of forecasts can be useful for groundwater management on a daily basis, for example in the scope of predicting groundwater level threshold values exceedance based on actual weather forecast, which are only reliable for short time windows.

Furthermore, the simulation mode has been tested: only the initial measured groundwater level values are fed into the network and the forecast is performed using the closed loop network for the whole time window considered. This kind of forecasts could be useful when investigating the potential effect of extreme seasons (e.g. a dry summer or a wet winter) or in the scope of climate change studies.

The most common workflow is to develop and train the network in the open loop structure and then use the closed loop structure in a second phase to perform the multi-step ahead forecasts (Beale et al., 2016), which make the training more effective and fast (Wunsch et al., 2018).

In this work each network has been trained in the open-loop structure, and then applied on the training set in the simulation mode with the closed loop structure. The choice of the hyper parameter (i.e. the parameter that are not learned by the training algorithm but must be chosen by the developer) has been based on the comparison of the SSE on the training set with the simulation mode.

The choice of the number of neuron and hidden layers has been performed through a grid-search. Particularly 10 to 40 neurons where tested for the first hidden layer, then at each step a second layer was added with an increasing number of neurons from 1 to 10 (Fig. S3 - S5).

The choice of the hyper parameters has been performed working with the pseudo-random generator seed number 1. Crane (2018) highlights that to avoid the problem of different pseudo-random generators it is advisable to report the results of neural network as a distribution of results obtained from a range of different seeds. To investigate the effect of the random initialization of weights and biases, after choosing the hyperparameters (i.e. number of neurons, activation functions, etc.), the selected models have been trained on five different runs, variating the seed of the random number generator. Results are reported as mean and standard deviation over these different runs.

(a) Levenberg-Marquardt training algorithm

Levenberg-Marquardt algorithm is an optimization method designed to work with loss functions with the form of a sum of squared error. Similarly to the Quasi-Newton method, it was developed to approach a second-order optimization problem by approximating the Hessian matrix H as follows:

$$\mathrm{H} = \mathrm{J}^T\mathrm{J} \tag{4}$$

So that the gradient G can be computed as

$$\mathrm{G} = \mathrm{J}^T\mathrm{e} \tag{5}$$

Where J is the Jacobian matrix containing the first derivative of the error function with respect to the biases and weights of the network and e is the vector of network errors. Backpropagation is used to calculate the Jacobian matrix J of the error function with respect to the weights and biases. In the present study biases and weights are initialized as random numbers within -1 and 1. At each k iteration of the training process, the gradient and the approximation of the Hessian matrix are calculated and consequentially the biases and weights $x_k$ of the network are updated according to the Levenberg-Marquardt method:

$$x_k = x_{k-1} - (\mathrm{J}^T\mathrm{J} + \mu\mathrm{I})^{-1}\mathrm{J}^T\mathrm{e} \tag{6}$$

The scalar μ is the Levenberg-Marquardt parameter, which has a role similar to a learning rate. When μ is zero Eqn 6 is a Quasi-Newton method; on the other hand, when μ is large Eqn 6 becomes a gradient descent with a small learning rate. The Newton's method is more accurate and faster close to a minimum of the error function. Therefore, the parameter μ is decreased by a certain factor after each iteration resulting in a decrease of the error function while it is increased by a certain factor when tentative step would increase the performance function. This means that the value of μ is increased until the change applied to weights and biases would result in a reduced performance value. Training stops when the maximum μ has been reached or when the minimum gradient or a performance goal (i.e. a minimum value of SSE) have been achieved, alternatively if a maximum number of iteration occurred.

After testing different initial values of the parameter μ, from 0.0001 to 0.01, it has been chosen to set it to 0.001 for the network trained with the Levenberg-Marquardt algorithm and early

stopping procedure, and to 0.005 for the networks trained with Levenberg-Marquardt algorithm and Bayesian Regularization. These values gave the best performance, as it can be shown by Fig. S6. The minimum gradient was set to 1e-7, maximum number of iterations to 10000, the performance goal to 0.

(b) Early stopping

Early stopping is a widely used technique to foster network generalization, working on the stopping criteria of the training algorithm. It consists in separating a subset of data from the training set and using them as a validation set. At each iteration of the training algorithm, the error function is calculated on both the training and the validation sets: weights and biases are updated based on the error on the training set, while the error function on the validation set is compared with those obtained in the previous iterations. The learning stops if the error function on the validation set increases for 10 iterations. This method prevents the network to overfit the training data which should lead to a better performance on different data. To the purpose of training the models with the early stopping procedure the training set (Fig. 2) was divided randomly into two subsets: 70% of the data were used as a proper training set, while the remaining 30% as validation set. Training and validation data must possess the same statistical properties (Maier and Dandy, 2000). To this purpose mean and standard deviations of the two subsets have been analysed, verifying their similarity: the means of the selected training and validation data are respectively 57.63 and 57.62, while the standard deviations are respectively 1.22 and 1.24.

The percentage of training and validation set influences the accuracy of the networks. Therefore, the training has been performed with several different percentages and the resulting networks were tested on the simulation mode. The best accuracy was reached with 70% for training and 30% for validation. Performances on the simulation mode of the network trained with different percentages of training and validation data are shown in Fig. S7.

(c) Bayesian regularization

Another method to deflect overfitting is called regularization. It consists in customizing the error function by adding an additional term that penalizes weights and biases with higher absolute values. In this study, regularization was implemented in the form of sum of the weights and biases, resulting in the following error function:

$$E = \alpha\, SSE + \beta \sum_{i=1}^{W} w_i \quad (7)$$

Where $w_i$ are the weights of the network and W is the total number of weights. If the scalar β is zero, the error function has the form of a normal sum of square error. By increasing the value of β with respect to α, the relevance of the regularization term increases. Using this modified error function causes the network to have smaller weights and biases, and this determines a smoother network response and reduces the chances of overfitting

Usually α and β belong to the family of the hyperparameters. Several approaches have been studied to choose the hyperparameters and in this study the Bayesian approach is used to optimize α and β (MacKay, 1992). The Bayesian approach considers weights and biases of the network as random variables with zero-mean gaussian prior distributions. The regularization parameters are related to the unknown variances associated with these distributions. These parameters are then optimized at each iteration according to the Bayes rule; in this way the objective function parameters are changing at each iteration of the training phase, leading to a changing objective function. For a more detailed description of the Bayesian regularization applied to neural network refer to (Foresee and Hagan, 1997).

To combine the Bayesian regularization and the early stopping procedure, training data were split into training and validation sets. The error function at Eqn. 7 is used during the training phase, updating α and β at each iterations using only the training data, and the iterations stop when the error on the validation set stops decreasing and starts increasing for ten consecutive iterations.

#### *3.3.3.7 Model comparison*

In order to compare the effectiveness of the developed models several statistical measures can be used to describe the error associated to the model output. After each model has been trained, its performance can be compared in terms of statistical measures of accuracy. In this study, SSE (Eqn. 3), the Mean Square Error (MSE), and the Root Mean Squared Error (RMSE) have been taken into account to compare the efficiency of the models as predictive tools. SSE is a measure of the residual variance, while MSE and RMSE are a measure of the standard deviation of the residuals.

$$MSE = \frac{SSE}{n} \tag{8}$$

$$RMSE = \sqrt{\frac{SSE}{n}} \tag{9}$$

Where $y_i$ is the i-th target value of groundwater level, $\hat{y}_i$ is the i-th groundwater level value calculated by the model; n is the total amount of data in the sequence. The smaller the values of SSE, MSE and RMSE, the better the model performance.

A variety of performance criteria can help evaluating the performance of the models, as different criteria can generally emphasize different aspects of the models predictive power (Maier et al., 2010). Here, Nash–Sutcliffe efficiency (NSE), Kling-Gupta efficiency (KGE) and Akaike Information Criterion (AIC) have been calculated:

$$NSE = 1 - \frac{MSE}{\sigma_o^2} \tag{10}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \quad with\ \alpha = \frac{\sigma_s}{\sigma_o}\ ,\ \mu = \frac{\mu_s}{\mu_o} \tag{11}$$

$$AIC = n\ log\left(\frac{SSE}{n}\right) + 2p \tag{12}$$

Where $\sigma_o$ and $\mu_o$ are respectively the mean and the standard deviation of the observed values while $\sigma_s$ and $\mu_s$ are the mean and standard deviation of the simulated values. The scalar p is the number of parameters of the model that are calculated during the training, which means the number of coefficients for the linear models and the number of weights and biases for the neural networks.

The NSE is one of the most widely used criteria for assessment of the hydrological models' performance, providing a measure of ability of the model to predict observed values (Shoaib et al., 2016). KGE is a reformulation of NSE, based on the Euclidian distance of the three NSE components from the ideal point, which overcomes the issues related to NSE (Gupta et al., 2009). Information criteria, such as the AIC, consider also model complexity, in addition to model error. Consequently, they have the potential to result in more parsimonious models (Maier et al., 2010).

All the metrics have been calculated on non-standardized data, so that the values of the residuals are expressed in meters.

The implemented models were tested on the one-day-ahead forecasting, as well as on multi-day-ahead forecasting. Moreover, the models were tested in the simulation mode: only the initial measured values of groundwater level are fed into the network and then the closed loop structure is used so that the model reproduces a complete multi-year scenario.

### 3.3.4 Results and discussion

The results of the network size search show that the best performing network trained with the early stopping method has 24 neurons in the first hidden layer and 4 in the second one (Fig. S3). The best performing network among those trained with both early stopping and Bayesian regularization has 30 neurons in the first hidden layer and 9 in the second one (Fig. S4), while the network trained with Bayesian regularization that outperformed the others is composed by 15 neurons in the first hidden layer and 1 in the second one (Fig. S5). The final setup of the selected neural networks are presented in Table 1

**Table 1 – Hyperparameter setup of the selected neural networks**

| | Early stopping | Bayesian Regularization | Early stopping + Bayesian Regularization |
|---|---|---|---|
| Initial μ | 0.001 | 0.005 | 0.005 |
| Max num of iteration | 10000 | 10000 | 10000 |
| Activation function in the hidden layers | tansig | tansig | tansig |
| Activation function in the output layer | linear | linear | linear |
| Hidden layer number | 2 | 2 | 2 |
| Hidden layers sizes | 24 - 4 | 15 - 1 | 30 - 9 |
| Num of iteration required (run with seed #1) | 27 | 777 | 27 |

Table 2 shows the performance statistics of the elaborated models on the training and the test set based on the different forecasting horizons, in terms of average performances over different runs and their standard deviations. The neural network trained with Bayesian regularization significantly outperforms the other models on both short and long-term forecasting on the training set. On the other hand, in the simulation mode, the model trained with the early stopping procedure shows the highest performance on the training set.

## Table 2 – Performance statistics of the forecasting models

| | | Training set | | | | | | | Test set | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Early stopping + Bayesian regularization | | Early stopping | | Bayesian Regularization | | ARX | Early stopping + Bayesian regularization | | Early stopping | | Bayesian Regularization | | ARX |
| | | **MEAN** | STDEV | **MEAN** | STDEV | **MEAN** | STDEV | | **MEAN** | STDEV | **MEAN** | STDEV | **MEAN** | STDEV | |
| 1 day | MSE | 0.002 | 1.39E-04 | 0.002 | 3.71E-04 | **0.001** | 3.09E-04 | 0.003 | 0.019 | 8.51E-04 | 0.060 | 5.10E-02 | 0.026 | 1.09E-03 | **0.010** |
| | RMSE | 0.045 | 1.54E-03 | 0.048 | 3.93E-03 | **0.026** | 5.89E-03 | 0.056 | 0.137 | 3.12E-03 | 0.229 | 8.83E-02 | 0.162 | 3.37E-03 | **0.099** |
| | SSE | 2.553 | 1.71E-01 | 2.880 | 4.59E-01 | **0.855** | 3.82E-01 | 3.868 | 11.364 | 5.13E-01 | 36.209 | 3.07E+01 | 15.788 | 6.55E-01 | **5.961** |
| | NSE | 0.999 | 9.12E-05 | 0.998 | 2.44E-04 | **1.000** | 2.03E-04 | 0.998 | 0.983 | 7.65E-04 | 0.946 | 4.58E-02 | 0.976 | 9.75E-04 | **0.992** |
| | KGE | 0.998 | 1.97E-03 | 0.997 | 2.36E-03 | **0.999** | 3.98E-04 | 0.999 | 0.944 | 3.31E-03 | 0.912 | 3.58E-02 | 0.963 | 1.36E-02 | **0.990** |
| | AIC | -4666.235 | 8.44E+01 | -5379.197 | 2.06E+02 | **-7887.375** | 5.70E+02 | -7117.538 | 582.557 | 2.76E+01 | 277.811 | 3.95E+02 | -961.065 | 2.52E+01 | **-2762.074** |
| 5 days | MSE | 0.014 | 1.52E-03 | 0.014 | 3.54E-03 | **0.004** | 1.62E-03 | 0.037 | 0.143 | 8.00E-03 | 0.274 | 1.87E-01 | 0.161 | 3.08E-02 | **0.037** |
| | RMSE | 0.118 | 6.54E-03 | 0.117 | 1.55E-02 | **0.063** | 1.22E-02 | 0.192 | 0.379 | 1.05E-02 | 0.499 | 1.55E-01 | 0.399 | 3.78E-02 | **0.192** |
| | SSE | 17.089 | 1.87E+00 | 17.199 | 4.36E+00 | **5.040** | 1.99E+00 | 45.691 | 85.743 | 4.78E+00 | 163.578 | 1.12E+02 | 96.228 | 1.84E+01 | **45.691** |
| | NSE | 0.991 | 9.92E-04 | 0.991 | 2.32E-03 | **0.997** | 1.06E-03 | 0.976 | 0.868 | 7.38E-03 | 0.748 | 1.72E-01 | 0.852 | 2.84E-02 | **0.932** |
| | KGE | 0.988 | 9.44E-03 | 0.983 | 1.57E-02 | **0.996** | 1.67E-03 | 0.983 | 0.814 | 1.15E-02 | 0.766 | 8.20E-02 | 0.861 | 5.02E-02 | **0.956** |
| | AIC | -2294.873 | 1.39E+02 | -3171.171 | 3.38E+02 | **-5622.914** | 4.64E+02 | -4041.015 | 1815.618 | 3.31E+01 | 1251.714 | 3.24E+02 | 132.949 | 1.12E+02 | **-1473.655** |
| 7 days | MSE | 0.018 | 2.40E-03 | 0.018 | 4.39E-03 | **0.005** | 2.23E-03 | 0.051 | 0.173 | 9.89E-03 | 0.308 | 1.91E-01 | 0.192 | 3.84E-02 | **0.111** |
| | RMSE | 0.134 | 8.99E-03 | 0.132 | 1.70E-02 | **0.072** | 1.44E-02 | 0.225 | 0.415 | 1.19E-02 | 0.533 | 1.52E-01 | 0.436 | 4.25E-02 | **0.333** |
| | SSE | 22.263 | 2.95E+00 | 21.777 | 5.40E+00 | **6.665** | 2.75E+00 | 63.132 | 102.955 | 5.90E+00 | 183.304 | 1.14E+02 | 114.338 | 2.29E+01 | **67.784** |
| | NSE | 0.988 | 1.57E-03 | 0.988 | 2.87E-03 | **0.996** | 1.46E-03 | 0.966 | 0.840 | 9.18E-03 | 0.715 | 1.77E-01 | 0.822 | 3.57E-02 | **0.906** |
| | KGE | 0.986 | 1.09E-02 | 0.979 | 1.96E-02 | **0.995** | 1.88E-03 | 0.976 | 0.800 | 1.26E-02 | 0.748 | 8.51E-02 | 0.843 | 5.67E-02 | **0.942** |
| | AIC | -1962.644 | 1.66E+02 | -2867.972 | 3.26E+02 | **-5268.758** | 4.64E+02 | -3638.148 | 1930.481 | 3.41E+01 | 1339.297 | 3.03E+02 | 240.888 | 1.13E+02 | **-1271.851** |

**Table 2 – Performance statistics of the forecasting models**

| | | Training set | | | | | | | Test set | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Early stopping + Bayesian regularization | | Early stopping | | Bayesian Regularization | | ARX | Early stopping + Bayesian regularization | | Early stopping | | Bayesian Regularization | | ARX |
| | | **MEAN** | STDEV | **MEAN** | STDEV | **MEAN** | STDEV | | **MEAN** | STDEV | **MEAN** | STDEV | **MEAN** | STDEV | |
| | MSE | 0.037 | 7.30E-03 | 0.033 | 9.05E-03 | **0.009** | 2.12E-03 | 0.091 | 0.242 | 2.38E-02 | 0.319 | 9.65E-02 | 0.223 | 3.52E-02 | **0.200** |
| | RMSE | 0.190 | 1.88E-02 | 0.180 | 2.57E-02 | **0.095** | 1.06E-02 | 0.302 | 0.491 | 2.49E-02 | 0.559 | 8.40E-02 | 0.471 | 3.74E-02 | **0.447** |
| 15 days | SSE | 44.655 | 8.91E+00 | 40.578 | 1.11E+01 | **11.204** | 2.59E+00 | 113.909 | 142.112 | 1.40E+01 | 187.682 | 5.67E+01 | 131.065 | 2.07E+01 | **122.390** |
| | NSE | 0.976 | 4.74E-03 | 0.978 | 5.88E-03 | **0.994** | 1.37E-03 | 0.939 | 0.776 | 2.20E-02 | 0.705 | 8.93E-02 | 0.794 | 3.26E-02 | **0.830** |
| | KGE | 0.979 | 1.23E-02 | 0.968 | 2.80E-02 | **0.993** | 3.15E-03 | 0.953 | 0.785 | 1.99E-02 | 0.752 | 5.55E-02 | 0.824 | 4.77E-02 | **0.903** |
| | AIC | -1085.262 | 2.39E+02 | -2077.471 | 3.64E+02 | **-4521.471** | 2.63E+02 | -2902.802 | 2139.899 | 6.13E+01 | 1432.298 | 1.75E+02 | 345.978 | 9.36E+01 | **-909.637** |
| | MSE | 0.073 | 2.00E-02 | 0.061 | 2.02E-02 | **0.016** | 2.21E-03 | 0.133 | 0.283 | 4.32E-02 | 0.351 | 1.11E-01 | **0.253** | 4.92E-02 | 0.311 |
| | RMSE | 0.267 | 3.50E-02 | 0.243 | 4.51E-02 | **0.124** | 8.77E-03 | 0.365 | 0.530 | 4.30E-02 | 0.585 | 9.29E-02 | **0.500** | 5.19E-02 | 0.558 |
| 30 days | SSE | 87.569 | 2.41E+01 | 73.670 | 2.44E+01 | **18.756** | 2.66E+00 | 165.721 | 162.116 | 2.47E+01 | 201.274 | 6.33E+01 | **144.948** | 2.82E+01 | 190.511 |
| | NSE | 0.953 | 1.28E-02 | 0.961 | 1.30E-02 | **0.990** | 1.41E-03 | 0.912 | 0.744 | 3.91E-02 | 0.682 | 1.00E-01 | **0.771** | 4.45E-02 | 0.735 |
| | KGE | 0.963 | 1.23E-02 | 0.953 | 3.25E-02 | **0.991** | 5.01E-03 | 0.917 | 0.800 | 3.49E-02 | 0.783 | 4.60E-02 | 0.812 | 5.05E-02 | **0.867** |
| | AIC | -224.648 | 2.99E+02 | -1330.692 | 5.02E+02 | **-3797.143** | 1.69E+02 | -2435.663 | 2246.720 | 9.85E+01 | 1501.839 | 1.82E+02 | 435.575 | 1.26E+02 | **-638.386** |
| | MSE | 0.240 | 1.86E-01 | **0.132** | 5.61E-03 | 0.218 | 1.84E-01 | 0.214 | 0.809 | 3.23E-01 | 0.515 | 1.79E-01 | **0.278** | 5.03E-02 | 0.495 |
| | RMSE | 0.454 | 1.85E-01 | **0.364** | 7.66E-03 | 0.427 | 1.89E-01 | 0.463 | 0.878 | 1.94E-01 | 0.705 | 1.31E-01 | **0.525** | 5.00E-02 | 0.703 |
| sim | SSE | 296.779 | 2.30E+02 | **163.544** | 6.94E+00 | 269.289 | 2.28E+02 | 266.968 | 487.538 | 1.95E+02 | 310.361 | 1.08E+02 | **167.599** | 3.03E+01 | 303.113 |
| | NSE | 0.842 | 1.22E-01 | **0.913** | 3.69E-03 | 0.857 | 1.21E-01 | 0.858 | 0.274 | 2.90E-01 | 0.538 | 1.60E-01 | **0.750** | 4.52E-02 | 0.579 |
| | KGE | 0.885 | 6.11E-02 | **0.945** | 1.28E-02 | 0.886 | 1.01E-01 | 0.816 | 0.800 | 8.06E-02 | **0.863** | 3.69E-02 | 0.834 | 5.82E-02 | 0.848 |
| | AIC | 825.903 | 9.79E+02 | -370.965 | 5.17E+01 | -1111.157 | 1.10E+03 | **-1841.544** | 2786.955 | 2.98E+02 | 1686.511 | 2.38E+02 | 452.680 | 1.21E+02 | **-353.712** |

As regards the test set results two different situations can be distinguished: the short-term forecasting, up to 15 days ahead, and the long-term forecasting from 30 days ahead up to the simulation mode. In the first case, when the considered forecasting horizon is close to the number of time lags considered by the model (i.e. 10 days) the linear model outperforms the neural networks; particularly the performance of the linear model is significantly better when considering the 1 and 5 days ahead forecasting while the statistical measures are comparable on the 7 day ahead forecasting. For the longer scenarios, 30 days ahead and the simulation mode, the neural network trained with Bayesian regularization shows a higher accuracy compared to the other networks and to the linear model.

Fig. 4 shows the comparison of real groundwater levels used as test set, which presents hydrological extremes exceeding the range of the training set, with those obtained by the linear model and by the neural network trained with the Bayesian regularization. The results of the neural network are the average over the five different runs performed with different initial weight and biases, and the standard deviation is represented by the grey error bars. Fig. 4a to Fig. 4e show the results of respectively 5, 7, 15, 30 steps ahead forecasting while Fig. 4f shows the comparison on the simulation mode.

Fig. 4a shows that both the ARx and NNARx have a high accuracy on the 1 day ahead forecasting and that the results of the NNARx are robust, with a reduced uncertainty. Fig. 4b and Fig. 4c highlight a noisier behaviour of the NNARx. In these cases, the ARx model reproduces the two highest peaks, that are outside of the range of the training set, with a higher accuracy. The peaks reproduced by the ARx model appear shifted in time. NNARx, on the other hand, slightly underestimates those peaks but reaches its maximum on the same days as the real groundwater level. With regard to the 15-days-ahead forecasting results, Fig. 4d shows that the NNARx reproduces the peaks and the summer decrease with a higher accuracy, while the temporal shift of the results of the ARx model are more evident. For the 30-days-ahead forecasting and on the simulation mode (Fig. 4e, 4f), NNARx clearly outperforms the ARx model that underestimates more than one half of the considered time series and fails at reconstructing the summer decrease and the second considered winter.

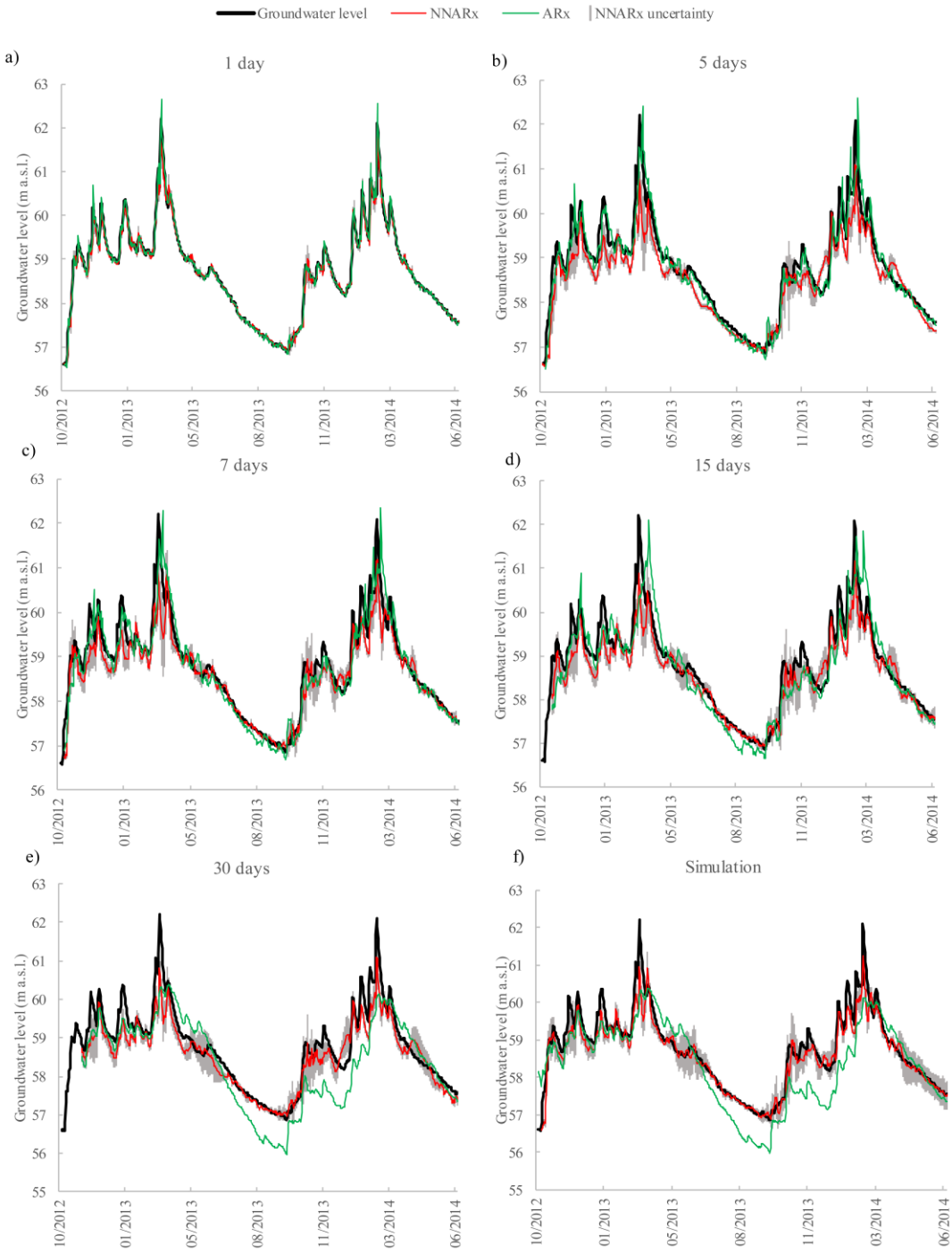



**Figure 4. Comparison of daily Groundwater levels and results of ARx and NNARx trained with Bayesian regularization on a) 1 day ahead, b) 5 days ahead, c) 7 days ahead, d) 15 days ahead, e) 30 days ahead and f) simulation mode (i.e. 2 years scenario).**

The uncertainty on the results of the NNARx varies in the considered time window. On the 30-days-ahead forecasting and the simulation mode, uncertainty appears to be higher especially at the beginning of the decreasing phase.

Table 2 shows that the performance of the neural network trained with early stopping is comparable with the one of the ARx model on the simulation mode, while the performance of the network trained with both early stopping and Bayesian regularization is lower. Particularly, in both cases the standard deviations of the results are higher as compared to the network trained with Bayesian regularization. This means that, even if single runs have higher performances than the ARx model (minimum SSE over the single runs on the simulation mode are 151.11 for early stopping and 178.06 for early stopping with Bayesian regularization), the averaged results over multiple runs show a lower performance. Fig. S8 shows the averaged results of these two NNARx models on the test set with the simulation mode. The NNARx trained with early stopping reproduces the groundwater level more correctly compared to the ARx model, apart from the spring 2014 when the ARx model outperforms the NNARx. The neural network trained with both early stopping and Bayesian regularization shows higher uncertainties and lower performance compared to the other neural networks and the linear model ARx.Compared to the previous ARx model (Zanotti et al., 2019), which considers only the precipitation as input variable, the performance of the current ARx model is increased due to the presence of the input variables temperature and time variable.

The residual analysis of the linear model ARx shows that the average of the residual is -3.8e-06 m, and that there is no autocorrelation among the residuals, and no cross-correlation between the residuals and the exogenous input variables (Fig. 5). To test the randomness of the residuals a runs test has been performed, which returns a decision for the null hypothesis that the values of the error vector come in a random order against the alternative hypothesis that they do not. The test is based on the number of consecutive values above or below the mean of the residuals vector. The resulting p-value of 0.56 indicates that the test does not reject the null hypothesis that the values in the residuals vector are in random order. The residual analysis results indicate that the error can be associated to a white noise, suggesting that the system can be successfully represented by a linear model.

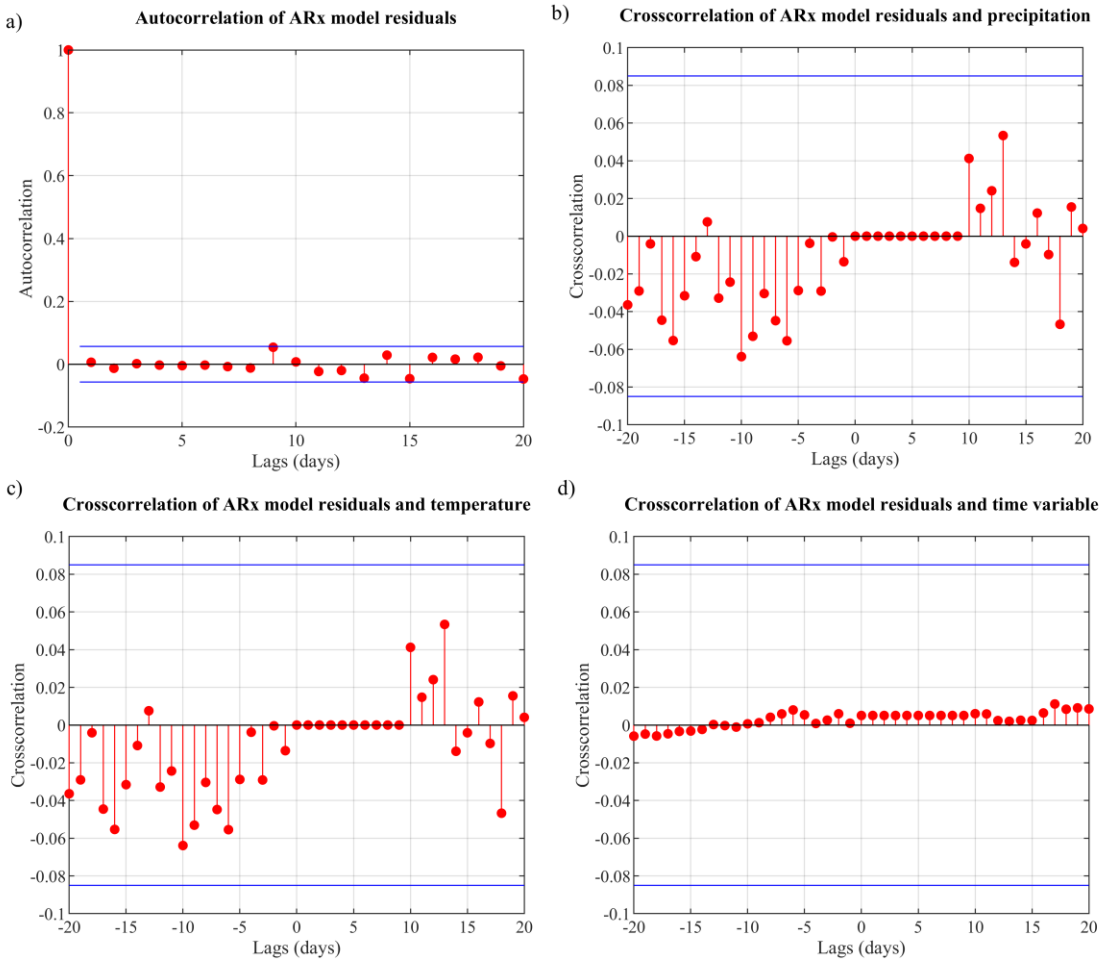



**Figure 5. Residual analysis of the ARx variable: a) Autocorrelation of the residuals, b) cross correlation between residuals and precipitation, c) cross correlation between residuals and temperature, d) cross correlation between residuals and time variable**

These results confirm that the linear models can be a suitable and easy-applicable tool for short-term groundwater level forecasting when the system can be considered linear, as their simplicity naturally prevents the risk of overfitting. This leads to a more reliable prediction even when hydrological conditions differ from those that prevailed during the training period. On the other hand, when longer scenarios are needed, for example in the scope of evaluating possible effects of dryer or rainier seasons, neural networks can offer a higher accuracy, especially if during the training phase proper precautions to avoid overfitting have been implemented.

The higher accuracy of the NNARx model on longer scenarios does not undermine the validity of the assumptions about the linearity of the considered system, which are confirmed by the results of the ARx model and its residual analysis. Indeed neural networks can solve both nonlinear and linear time series forecasting problems (Zhang, 2001). Furthermore, the fact that NNARx can reach a higher accuracy on long-term forecasts suggests that they could be a more suitable tool when studying long-term scenarios, for example in the scope of climate changes.

The ARx model reproduces more correctly the amplitude of the peaks on the short-term forecasting but the results appear shifted in time. This is due to the simple structure of the ARx model which relies with a strong dependency on the observed groundwater data leading to results that are shifted in time of a number of days close to the forecasting horizon (Ginocchi et al., 2016; Valipour et al., 2013), and this aspect must be taken into account when using the model for managing purposes. On the other hand neural networks, whose structure have been chosen based on their performance on the simulation mode on the training set, show a weaker dependency on the observed groundwater data and offer a higher accuracy even when working with their own output as groundwater level input.

In most cases, neural networks applied to groundwater level forecasting problems have only one hidden layer (Banerjee et al., 2009; Mohanty et al., 2010; Taormina et al., 2012; Yang et al., 2009); this study highlights that, in the considered case study, adding a second hidden layer can improve the network performance.

Among the different techniques to avoid overfitting considered in this study, Bayesian regularization shows a higher accuracy compared to early stopping when applied to a test set. Combining Bayesian regularization and early stopping decreases the accuracy of the model: Bayesian regularization in fact requires a higher number of iterations to converge while early stopping works by limiting the number of iterations and thus avoids the Bayesian regularization to reach a proper convergence (Table 1). The higher number of iterations required by the Bayesian regularization (Table 1) could become computationally and time consuming: in this case for example the chosen Bayesian regularization network (hidden layers sizes 15 - 1) took ca. 5 minutes to conclude the 777 iterations, while the biggest network tested (hidden layers sizes 40 - 10) took ca. 20 hours to conclude the 870 iterations using an Intel®

Core™ i7-7700HQ with RAM 16 GB. This means that if larger datasets are used (e.g. hourly data) developing and testing a wide range of networks could become more time consuming. On the other hand, this study shows that when it is possible to properly apply this technique it can be considered a reliable tool to prevent overfitting when implementing feed forward neural networks.

### 3.3.5 Conclusion

This work presents the comparison between linear models (ARx) and neural networks (NNARx) for groundwater level forecasting in a system where a linear relationship can be assumed between groundwater level and precipitation. For the neural networks, which are known for their tendency to overfit, a special focus has been made on the ability of the different models to avoid overfitting. Three different methods to avoid overfitting have been tested: early stopping, Bayesian regularization and the combination of both.

The main conclusions highlighted by the results of this work are:

1) Data driven models developed on the studied system have been proven to be effective tools to fulfil the need, highlighted by the local authorities, of models capable to deal with hydroclimatic extremes, which can support a more effective and sustainable water resource management.

2) Linear models can be a suitable and easy-applicable tool for short-term groundwater level forecasting when the system can be considered linear. Since the complexity of the model is reduced in the case of ARx, the risk of overfitting is also naturally reduced.

3) Neural networks are more reliable when longer scenarios are needed, for example in the scope of evaluating possible effects of hydroclimatic extremes (e.g. dryer or rainier seasons) if proper techniques to avoid overfitting are implemented during the training phase.

4) Bayesian regularization can be considered as a valid technique to prevent overfitting but coupling it with early stopping could decrease its performance. Therefore, if the dataset is too large to manage the amount of iterations required by Bayesian

regularization (e.g. hourly data) early stopping alone can be considered as an alternative. On the other hand, early stopping showed a higher instability so that, when averaged over different runs, the resulting performance was lower and comparable to the performance of linear models.

5) Even if usually one hidden layer is considered sufficient, these results highlight that adding a second hidden layer can increase the performance of the network.

The methodology here presented allows researchers to choose between linear and non linear forecasting models and avoid overfitting in systems where a linear recharge - groundwater level response can be assumed on the basis of the hydrogeological conceptual models. Groundwater level forecasting can be a useful tool to support decision makers but the developed model has to some extent be considered reliable also outside the range of the training set to be able to cope with hydroclimatic extremes.

### 3.3.6 References

Adamowski, J., Chan, H.F., 2011. A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol. 407, 28–40. https://doi.org/10.1016/j.jhydrol.2011.06.013

Adamowski, J., Karapataki, C., 2010. Comparison of Multivariate Regression and Artificial Neural Networks for Peak Urban Water-Demand Forecasting: Evaluation of Different ANN Learning Algorithms. J. Hydrol. Eng. 15, 729–743. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000245

Ahn, H., Salas, J.D., 1997. Groundwater head sampling based on stochastic analysis. Water Resour. Res. 33, 2769–2780.

Banerjee, P., Prasad, R.K., Singh, V.S., 2009. Forecasting of groundwater level in hard rock region using artificial neural network. Environ. Geol. 58, 1239–1246.

Beale, M.H., Hagan, M.T., Demuth, H.B., 2016. Neural Network Toolbox ™ User ' s Guide. The MathWorks, Inc.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B.,

Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20. https://doi.org/10.1016/j.envsoft.2012.09.011

Burrows, W., Doherty, J., 2015. Efficient Calibration/Uncertainty Analysis Using Paired Complex/Surrogate Models. Groundwater 53, 531-541. https://doi.org/10.1111/gwat.12257

Burrows, W., Doherty, J., 2016. Gradient-based model calibration with proxy-model assistance. J. Hydrol. 533, 114-127. https://doi.org/10.1016/j.jhydrol.2015.11.033

Candanedo, L.M., Feldheim, V., Deramaix, D., 2018. Reconstruction of the indoor temperature dataset of a house using data driven models for performance evaluation. Build. Environ. 138, 250–261. https://doi.org/10.1016/j.buildenv.2018.04.035

Chiaudani, A., Di Curzio, D., Palmucci, W., Pasculli, A., Polemio, M., Rusi, S., 2017. Statistical and Fractal Approaches on Long Time-Series to Surface-Water/Groundwater Relationship Assessment: A Central Italy Alluvial Plain Case Study. Water 9, 850. https://doi.org/10.3390/w9110850

Choubin, B., Malekian, A., 2017. Combined gamma and M-test-based ANN and ARIMA models for groundwater fluctuation forecasting in semiarid regions. Environ. Earth Sci. 76, 1–10. https://doi.org/10.1007/s12665-017-6870-8

Coppola, E., Poulton, M., Charles, E., Dustman, J., Szidarovszky, F., 2013. Application of Artificial Neural Networks to Complex Groundwater Management Problems. Water Resour. Manag. 27, 216–220. https://doi.org/10.1016/j.jhydrol.2016.01.076

Crane, M., 2018. Questionable Answers in Question Answering Research: Reproducibility and Variability of Published Results. Trans. Assoc. Comput. Linguist. 6, 241–252.

Daliakopoulos, I.N., Coulibaly, P., Tsanis, I.K., 2005. Groundwater level forecasting using artificial neural networks. J. Hydrol. 309, 229–240. https://doi.org/10.1016/j.jhydrol.2004.12.001

Dietterich, T., 1995. Overfitting and Undercomputing in Machine Learning. ACM Comput. Surv. 27.

Foresee, F.D., Hagan, M.T., 1997. Gauss-Newton approximation to Bayesian learning. Proc. Int. Jt. Conf. Neural Networks 1930–1935.

Gianni, G., Doherty, J., Brunner, P., 2019. Conceptualization and Calibration of Anisotropic Alluvial Systems: Pitfalls and Biases. Groundwater 57, 409-419. https://doi.org/10.1111/gwat.12802

Ginocchi, M., Crosta, G.F., Rotiroti, M., Bonomi, T., 2016. Analysis and prediction of groundwater level time series with Autoregressive Linear Models. Rend. Online Soc. Geol. It. 39, 109–112. https://doi.org/10.3301/ROL.2016.59

Grassi, S., Doveri, M., Ellero, A., Palmieri, F., Vaselli, L., Calvi, E., Trifirò, S., 2011. Studio dei sistemi termali di Montecatini e Monsummano Terme.

Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Harmel, R.D., Smith, P.K., Migliaccio, K.., Chaubey, I., Douglas-Mankin, K.R., Benham, B., Shukla, S., Munoz-Carpena, R., Robson, B.J., 2014. Evaluating, interpreting, and communicating performance of hydrologic/water quality models considering intended use: A review and recommendations. Environ. Model. Softw. 57, 40–51.

Hartmann, A., Mudarra, M., Andreo, B., Marin, A., Wagener, T., Lange, J., 2014. Modeling spatiotemporal impacts of hydroclimatic extremes on groundwater recharge at a Mediterranean karst aquifer. Water Resour. Res. 50, 6507–6521. https://doi.org/10.1002/2014WR015685.

Huang, G., Sun, J., Zhang, Y., Chen, Z., Liu, F., 2013. Impact of anthropogenic and natural processes on the evolution of groundwater chemistry in a rapidly urbanized coastal area, South China. Sci. Total Environ. 463–464, 209–221. https://doi.org/10.1016/j.scitotenv.2013.05.078

Izady, A., Davary, K., Alizadeh, A., Moghaddam Nia, A., Ziaei, A.N., Hasheminia, S.M., 2013. Application of NN-ARX Model to Predict Groundwater Levels in the Neishaboor Plain, Iran. Water Resour. Manag. 27, 4773–4794. https://doi.org/10.1007/s11269-013-0432-y

Juan, C., Genxu, W., Tianxu, M., 2015. Simulation and prediction of suprapermafrost groundwater level variation in response to climate change using a neural network model. J. Hydrol. 529, 1211–1220. https://doi.org/10.1016/j.jhydrol.2015.09.038

Kang, G., Li, J., Tao, D., 2018. Shakeout: A New Approach to Regularized Deep Neural Network Training. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1245–1258. https://doi.org/10.1109/TPAMI.2017.2701831

Khalil, B., Broda, S., Adamowski, J., Ozga-Zielinski, B., Donohoe, A., 2015. Short-term forecasting of groundwater levels under conditions of mine-tailings recharge using wavelet ensemble neural network models. Hydrogeol. J. 23, 121–141. https://doi.org/10.1007/s10040-014-1204-3

LeCun, Y.A., Bottou, L., Orr, G.B., Müller, K.-R., 2012. Efficient BackProp, in: Montavon, G., Orr, Geneviève B, Müller, K.-R. (Eds.), Neural Networks: Tricks of the Trade: Second Edition. Springer, Berlin, Heidelberg, pp. 9–48. https://doi.org/10.1007/978-3-642-35289-8_3

Lee, S.-I., Lee, S.-K., Hamm, S.-Y., 2009. A Model for Groundwater Time-series from the Well Field of Riverbank Filtration. J. Korea Water Resour. Assoc. 42, 673–680. https://doi.org/10.3741/JKWRA.2009.42.8.673

Lu, T., Lü, X., Kibert, C., 2015. A hybrid numerical-neural-network model for building simulation : A case study for the simulation of unheated and uncooled indoor temperature. Energy Build. 86, 723–734. https://doi.org/10.1016/j.enbuild.2014.10.024

MacKay, D.J.C., 1992. Bayesian Interpolation. Neural Comput. 4, 415–447. https://doi.org/10.1162/neco.1992.4.3.415

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables : a review of modelling issues and applications 15, 101–124.

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 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

Maskey, S., Dibike, Y.B., Jonoski, A., Solomatine, D., 2000. Groundwater model approximation with artificial neural network for selecting optimum pumping strategy for plume removal. Int. Inst. Infrastructural, Hydraul. Environ. Eng. Delft, Netherlands.

Mohanty, S., Jha, M.K., Kumar, A., Sudheer, K.P., 2010. Artificial neural network modeling for groundwater level forecasting in a river island of eastern India. Water Resour. Manag. 24, 1845–1865. https://doi.org/10.1007/s11269-009-9527-x

Nayak, P.C., Rao, Y.R.S., Sudheer, K.P., 2006. Groundwater level forecasting in a shallow aquifer using artificial neural network approach. Water Resour. Manag. 20, 77–90.

Nikolos, I.K., Stergiadi, M., Papadopoulou, M.P., Karatzas, G.P., 2008. Artificial neural networks as an alternative approach to groundwater numerical modelling and environmental design. Hydrol. Process. 22, 3337–3348.

Ozbalta, T.G., Sezer, A., Yildiz, Y., 2012. Indoor and Built Models for Prediction of Daily Mean Indoor Temperature and Relative Humidity : Education Building in Izmir , Turkey. Indoor build Environ. 21, 772–781. https://doi.org/10.1177/1420326X11422163

Rosin, P.L., Fierens, F., 1995. Improving Neural Network Generalisation. Igarss'95 2, 1255–1257. https://doi.org/10.1109/IGARSS.1995.521718

Rumelhart, D.E., McClelland, J.L., 1986. Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations. MIT Press, Cambridge, MA, USA.

Shirmohammadi, B., Vafakhah, M., Moosavi, V., Moghaddamnia, A., 2013. Application of Several Data-Driven Techniques for Predicting Groundwater Level. Water Resour. Manag. 27, 419–432. https://doi.org/10.1007/s11269-012-0194-y

Shoaib, M., Shamseldin, A.Y., Melville, B.W., Khan, M.M., 2016. A comparison between wavelet based static and dynamic neural network approaches for runoff prediction. J. Hydrol. 535, 211–225. https://doi.org/10.1016/j.jhydrol.2016.01.076

Sreekanth, P.D., Geethanjali, N., Sreedevi, P.D., Ahmed, S., Kumar, N.R., Jayanthi, P.D.K., 2009. Forecasting groundwater level using artificial neural networks. Curr. Sci. 933–939.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout:

A Simple Way to Prevent Neural Networks from Overfitting. J. Mach. Learn. Res. 15, 1929–1958. https://doi.org/10.1214/12-AOS1000

Stefania, G.A., Rotiroti, M., Fumagalli, L., Simonetto, F., Capodaglio, P., Zanotti, C., Bonomi, T., 2018. Modeling groundwater/surface-water interactions in an Alpine valley (the Aosta Plain, NW Italy): the effect of groundwater abstraction on surface-water resources. Hydrogeol. J. 26, 147-162. https://doi.org/10.1007/s10040-017-1633-x

Sudheer, K.S., Kasiviswanathan, K.P., 2017. Methods used for quantifying the prediction uncertainty of artificial neural network based hydrologic models. Stoch. Environ. Res. Risk Assess. 31, 1659–1670. https://doi.org/10.1007/s00477-016-1369-5

Taormina, R., Chau, K., Sethi, R., 2012. Artificial neural network simulation of hourly groundwater levels in a coastal aquifer system of the Venice lagoon. Eng. Appl. Artif. Intell. 25, 1670–1676. https://doi.org/10.1016/j.engappai.2012.02.009

Thomas, B., Soleimani-Mohseni, M., 2007. Artificial neural network models for indoor temperature prediction : investigations in two buildings. Neural Comput Applic 16, 81–89. https://doi.org/10.1007/s00521-006-0047-9

Tokar, A.S., Johnson, P.A., 1999. Rainfall-runoff modeling using artificial neural networks. J. Hydrol. Eng. 4, 232–239.

Tsanis, I.K., Coulibaly, P., Daliakopoulos, I.N., 2008. Improving groundwater level forecasting with a feedforward neural network and linearly regressed projected precipitation. J. Hydroinformatics

Valipour, M., Banihabib, M.E., Behbahani, S.M.R., 2013. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. J. Hydrol. 476, 433–441. https://doi.org/10.1016/j.jhydrol.2012.11.017

Wunsch, A., Liesch, T., Broda, S., 2018. Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX). J. Hydrol. 567, 743–758. https://doi.org/10.1016/j.jhydrol.2018.01.045

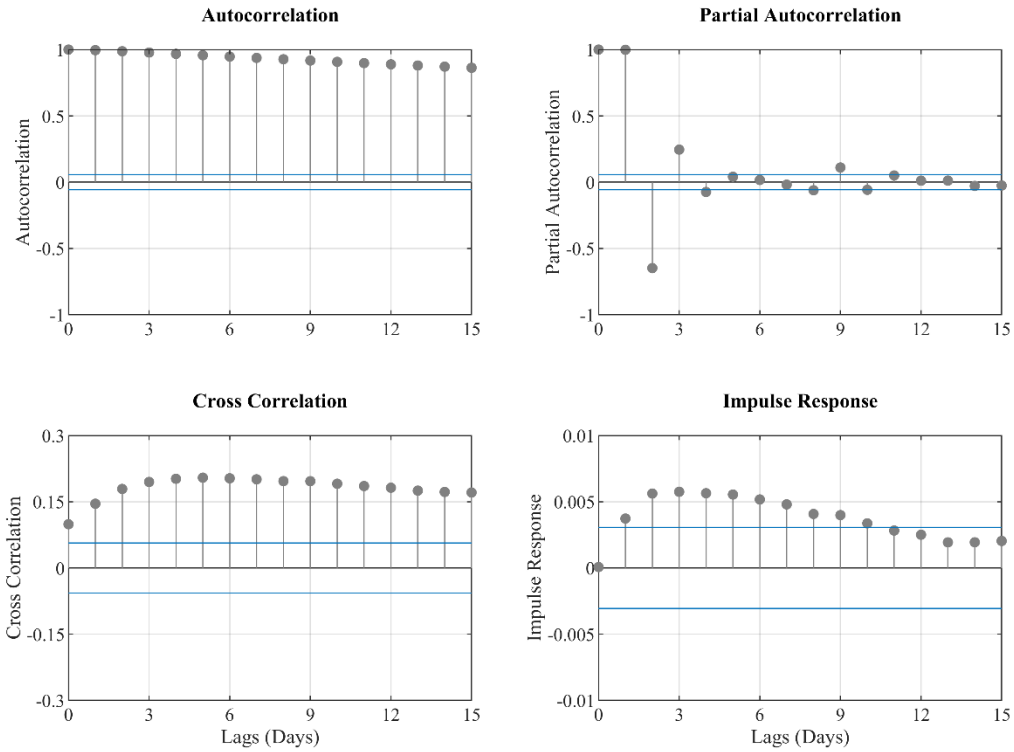
Yang, Z.P., Lu, W.X., Long, Y.Q., Li, P., 2009. Application and comparison of two prediction models for groundwater levels: A case study in Western Jilin Province, China. J. Arid Environ. 73, 487–492. https://doi.org/10.1016/j.jaridenv.2008.11.008

Yonaba, H., Anctil, F., Fortin, V., 2010. Comparing sigmoid transfer functions for neural network multistep ahead streamflow forecasting. J. Hydrol. Eng. 15, 275–283.

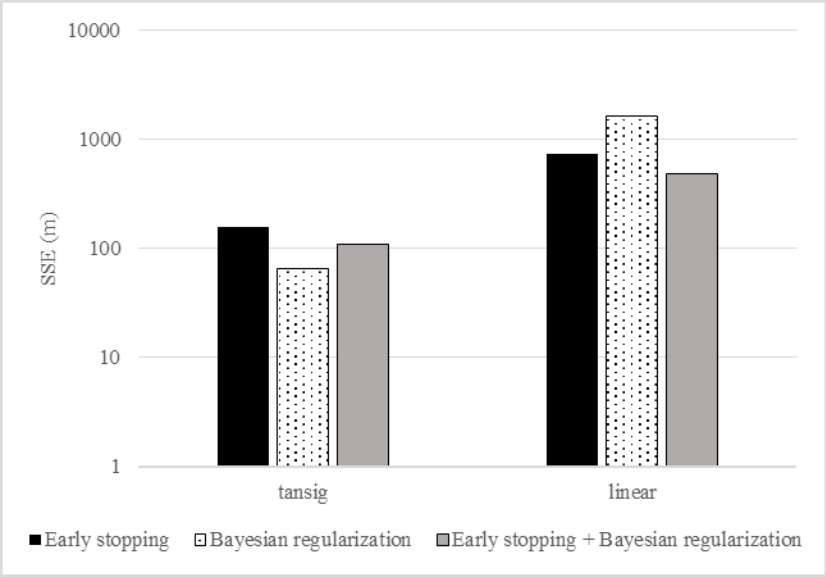
Zanotti, C., Bonomi, T., Nannucci, M., Rotiroti, M., 2019. Groundwater level forecasting using linear time series modeling: the case study of the thermal aquifer system of Monsummano Terme (central Italy). Rend. Online 47, 153–160. https://doi.org/10.3301/ROL.2019.27

Zhang, G.P., 2001. An investigation of neural networks for linear time-series forecasting. Comput. Oper. Res. 28, 1183}1202. https://doi.org/10.1016/S0305-0548(00)00033-2

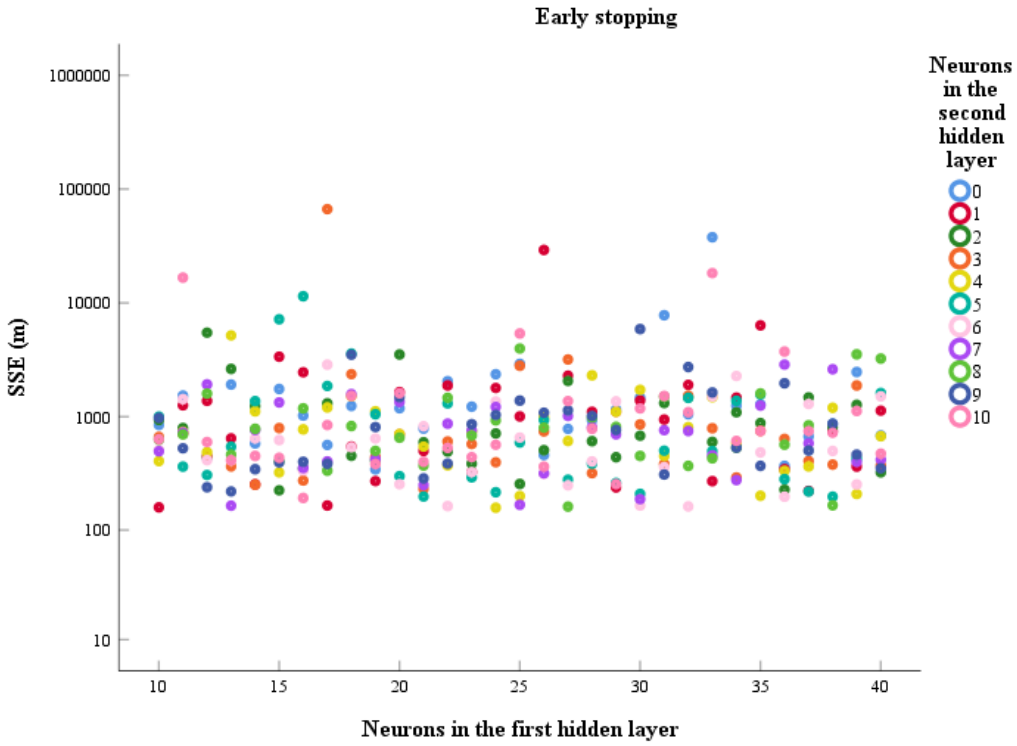
### 3.3.7 Supplementary material



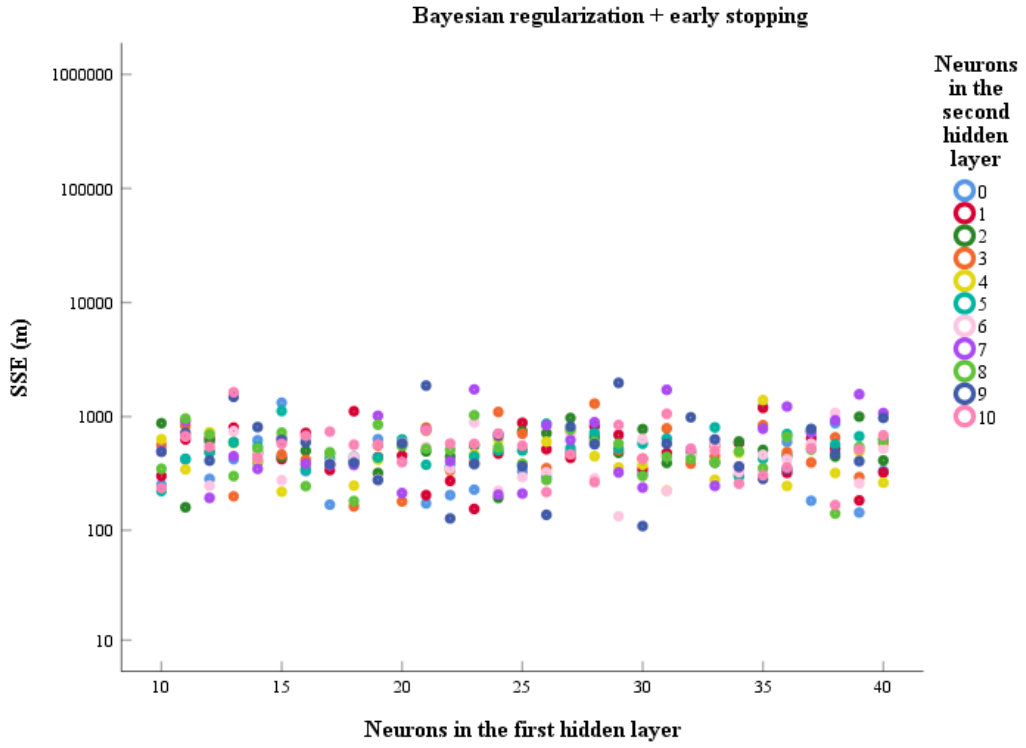

**S1 – Autocorrelation and Partial Autocorrelation of groundwater level on the top. Groundwater level Cross correlation and Impulse Response with precipitation on the bottom. Modified from Zanotti et al. (2019)**



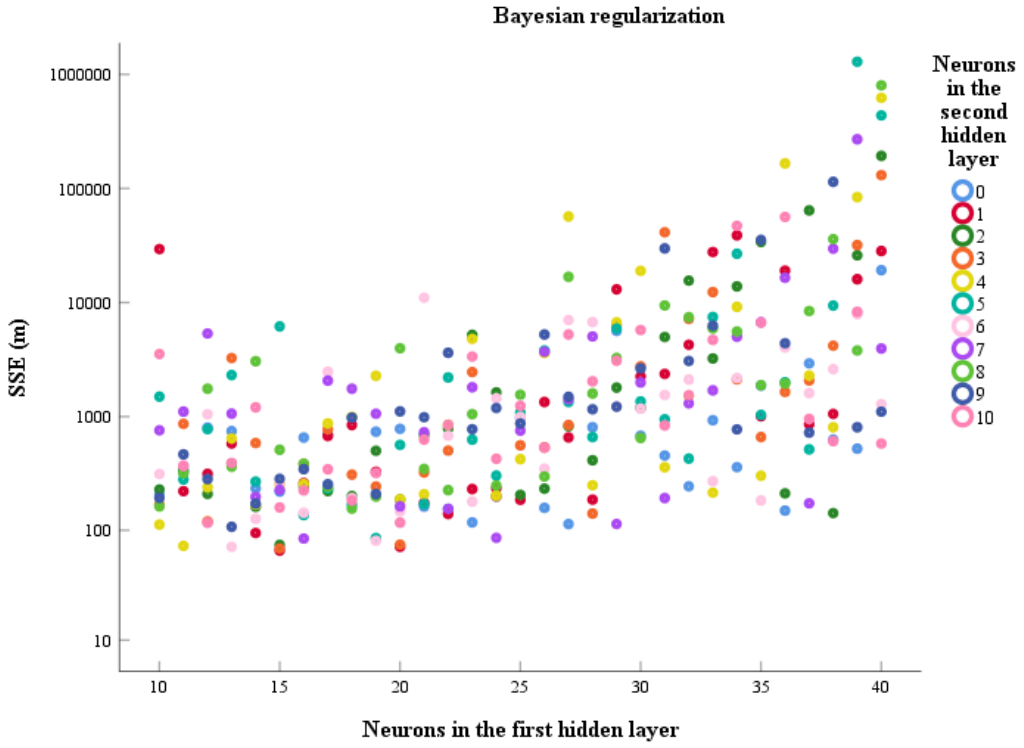

**S2 – Performance of the NNARx with tansig and linear activation functions in terms of SSE on the training set in the simulation mode**



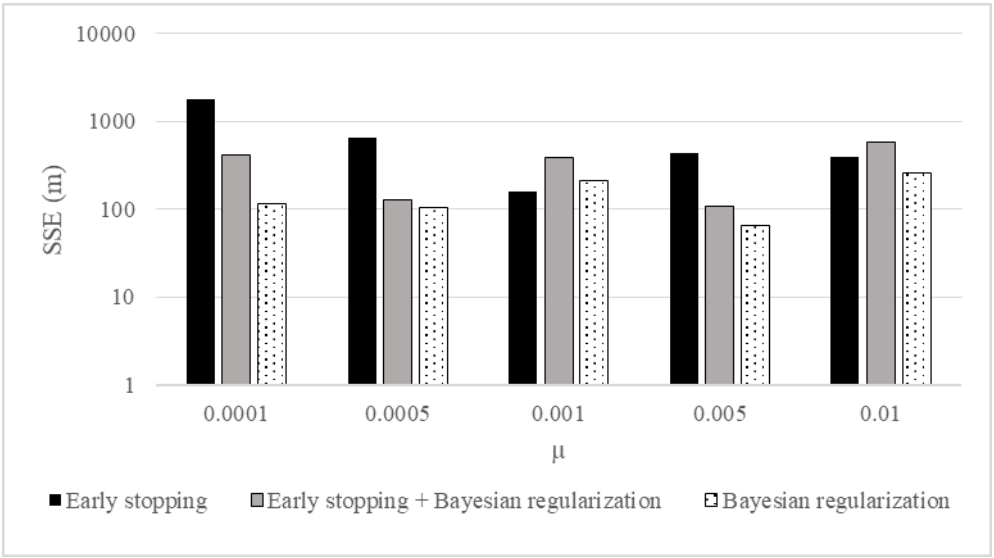

**S3 – Performance of the NNARx trained with early stopping in terms of SSE on the training set in the simulation mode**



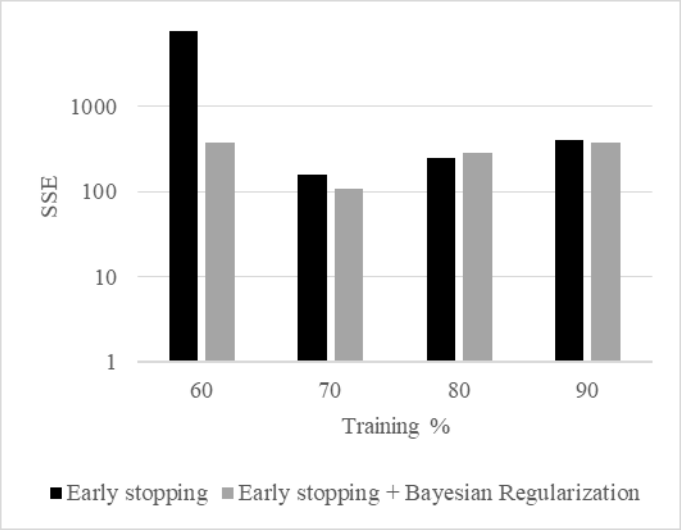

**S4 – Performance of the NNARx trained with Bayesian regularization and early stopping in terms of SSE on the training set in the simulation mode**



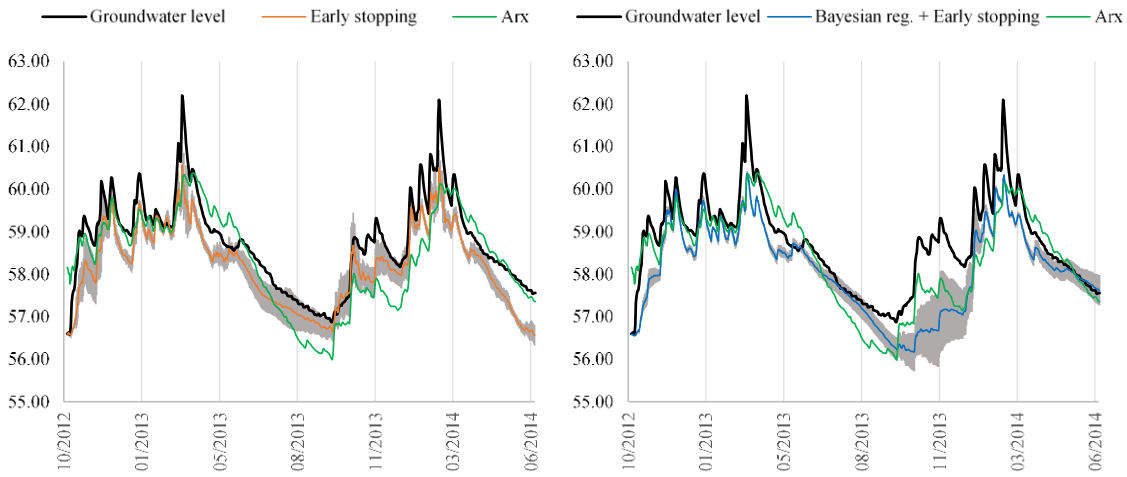

**S5 – Performance of the NNARx trained with Bayesian regularization in terms of SSE on the training set in the simulation mode**




**S6 – Performance of the NNARx trained with different initial μ values in terms of SSE on the training set in the simulation mode**




**S7 – SSE of network trained with different percentages of training and validation data on the whole training set (i.e. training + validation data) in the simulation mode.**




**S8 – Results of NNARx trained with a) Bayesian regularization and early stopping and b)early stopping on the test set, on the simulation mode.**

# Chapter 4
# **Conclusions**

In this PhD research several data-driven techniques were investigated and applied on real groundwater datasets. Two parallel workflows were followed, the first one related to the quality of the groundwater resource, while the second one to its quantity.

Tackling these two aspects of groundwater management led to facing two completely different problems from a data science perspective:

- The investigation of the groundwater quality dataset involved the analysis of multivariate data. In this case the dataset was composed of several samples, collected from groundwater and surface water bodies, described by several chemical variables. The main aim was to perform data mining, i.e. identifying the information hidden within the variability of the data. The goal of the implementation of multivariate analysis was to identify the different hidden factors controlling groups of variables and affecting differently groups of samples.
- The investigation of the groundwater quantity dataset involved the analysis of time series data. In this case the dataset was composed of time series of groundwater level and meteorological data collected in a fixed location. The main aim was to investigate the dynamic of the system and its "memory effect" and to develop forecasting tools that could be useful for running future scenarios.

In both cases the workflow consisted in two consecutive phases: a) a first exploratory analysis, aimed at understanding the potentiality of the datasets and how to settle the problem and its solution, and b) the implementation of a relatively innovative technique and the investigation and interpretation of the obtained results.

This work allowed for several considerations which are summarized in the following paragraphs.

## 4.1 Multivariate analysis

Overall findings reported in Chapter 2 confirm how hydrochemical datasets can be interpreted by multivariate statistical analysis. The primary benefit of using this method to interpret hydrochemical data compared to "traditional" techniques, such as Piper diagrams and bivariate plots, is that Multivariate statistics are self-sufficient and quantitative techniques, capable of extracting information from all available data by analysing all the considered variables and samples collectively and simultaneously. Results of multivariate statistical analysis have succeeded in improving the area's conceptual hydrological and hydrogeological model and summarising the impacts of groundwater / surface water interactions on water chemistry. The conceptual model enhancement, reached by means of multivariate analysis, can foster the development of the studied system's numerical flow modeling.

In this work, CA allowed for a meaningful classification of the sampling stations and supported the improvement of the monitoring network for the successive field campaign. Identifying groups of similar samples and outliers allowed for an efficient reduction of the monitoring network by avoiding redundant sampling points and points affected by phenomena which were not in the focus of the study.

FA and PMF are similar techniques, in terms of inputs and outputs. They both aim at reducing the dimensionality of a multivariate datasets identifying latent factors.

FA resulted to be an agile and well standardized method to quantify the hidden factors determining the variability of the analysed data. Although it presented some limitation it definitively demonstrated to be a nimble and objective method (the operator only has to decide which samples and variables to use) to explore multivariate datasets.

PMF allowed for a more detailed description of the system, revealing and characterizing more phenomena and features compared to FA. Furthermore, PMF allowed for considering the analytical uncertainty of the data and therefore for coping with missing and below detection limit data by associating to them a higher uncertainty. The positivity constraints of the PMF led to a more environmentally interpretable representation of the system compared to FA.

Although CA, FA and PMF were not designed as spatial analysis, the visualization and investigation of their results trough a GIS approach led to an easier interpretation and highlighted spatial pattern and feature of the system.

CA, FA and PMF revealed to be valuable tools to perform data mining on groundwater quality datasets, characterized by several variables. They can support the identification and quantification of the natural and anthropogenic sources affecting the systems, highlighting their spatial variability. Therefore, multivariate analysis could be considered as a useful tool to perform water quality characterization and to foster the development of the conceptual model in complex hydro-systems where different water resources are involved (e.g. groundwater, lake, rivers and springs).

## 4.2 Time series analysis

Results reported in chapter 3 highlight how timeseries analysis can be a valuable tool to explore and model groundwater level data over time. As opposite to groundwater numerical flow models, time series analysis does not require any information about the geology of the aquifer and its features.

In general, forecasts obtained with data driven techniques are more accurate within the limits of the observed data; in the present work a special focus was made on the application of the models on a test set where the groundwater level exceeds the range of the training set, leading to a more accurate quantification of the model accuracy outside the range of its training set.

In this work, univariate and bivariate analysis such as ACF, PACF, cross correlation and impulse response, revealed to be of use in performing exploratory analysis. Their results indeed can constitute a meaningful and solid basis for choosing the forecasting model type and orders.

Furthermore, the application of these techniques on the residuals of the forecasting models led to an in-depth examination of error terms. Beside validating the models with a separate test set, the residual analysis appeared to be a relevant aspect in the choice and interpretation of the models as it allows to understand whether the error still contains relevant information, a condition that should be avoided.

Linear models such as AR and ARx demonstrated to be suitable and easy-applicable tools for short-term groundwater level forecasting in systems that could be considered as linear. To develop these models the only choice that has to be done by the operator is the number of considered time lags. This, together with the low complexity given by the reduced number of coefficients, naturally prevents the risk of overfitting.

Neural networks demonstrated to be more reliable when working with longer scenarios. Although their implementation requires more effort as compared to linear models, their accuracy on long-term forecasting makes them potential valuable tools of working in the scope of extreme seasons or climate changes.

The risk of overfitting when working with neural network is well known. In this work Bayesian regularization demonstrated to be a valid technique to prevent overfitting.

## 4.3 General conclusions

Summarizing, the research work and the scientific results of this PhD led to the following considerations:

- In a landscape of increasing availability of environmental data, data science can actually become a useful instrument to enhance the understanding of the groundwater systems and to promote a more sustainable and efficient use of the groundwater resource.
- Data-driven techniques can support the development of the conceptual model of a system highlighting natural and anthropogenic sources and the relationship between groundwater level and precipitation.
- Although data-driven techniques rely on purely mathematical assumptions, when applied to hydrogeological datasets the hydrogeological competences can not be

neglected. Errors and noise in the data can potentially bring the results of the analysis toward non-realistic results. Therefore, only the joint effort of data science and hydrogeology can actually lead to an effective exploitation of the data potential.

- Data-driven techniques can only bring to light information that is actually gathered within the data. Therefore, the quality and the size of the considered dataset becomes a crucial aspect of the data science application to environmental data which are naturally affected by a higher uncertainty and variability if compared to other research fields.

# Appendix A
# Publications

During this PhD, the following activities were performed:

- Contribution for field sampling and data analysis to the project "Lake, stream and groundwater modeling to manage water quantity and quality in the system of Lake Iseo - Oglio river.", Cariplo grant n. 2014-1282. The project aimed at a) identifying the main hydrochemical features of either groundwater and surface water (lake, river and springs) and the processes that influence (or govern) them; b) to understand the relations (if any) between the chemistry of groundwater and surface water; c) developing a numerical flow model of the groundwater/surface water system. Publications: [5, 7, 8, 11, 12, 13, 15, 20, 21]
- Collaboration with Regione Toscana for the analysis of groundwater level data of Monsummano Terme (PT) and the development of a forecasting tool. Publications: [6, 14]
- Contribution to the collaboration with ARPA Valle d'Aosta, particularly in the scope of the work "Determination of trigger levels for groundwater quality in landfills located in historically human-impacted areas". This work presents a methodology for calculating trigger levels for groundwater quality in landfills located in areas where historical contaminations have deteriorated groundwater quality prior to their construction. This method is based on multivariate statistical analysis and involves 4 steps: (a) implementation of the conceptual model, (b) landfill monitoring data collection, (c) hydrochemical data clustering and (d) calculation of the trigger levels.

The proposed methodology was applied on a case study in northern Italy, where a currently used lined landfill is located downstream of an old unlined landfill and others old unmapped waste deposits. The developed conceptual model stated that groundwater quality deterioration observed downstream of the lined landfill is due to a degrading leachate plume fed by the upgradient unlined landfill. The methodology led to the determination of two trigger levels for COD and NH4-N, the former for a zone representing the background hydrochemistry (28 and 9 mg/L for COD and NH4-N, respectively), the latter for the zone impacted by the degrading leachate plume from the upgradient unlined landfill (89 and 83 mg/L for COD and NH4-N, respectively). Publications: [2, 3, 4, 9, 17]

- Contribution to the study about "Chloride increase in the deep south-Alpine lake Iseo (Northern Italy): load quantification and source identification with a watershed approach". In this study, through the analysis of a long-term dataset (1993-2017), the trend of chloride was analyzed together with the land use and population dynamic evolution in the lake mixed-land-use watershed with the aim to understand the possible relationship among them. Publications: [16, 18, 19]
- Contribution to the scientific support to Regione Lombardia for the determination of Natural Background Level (NBL) of groundwater. This work involves the analysis of groundwater quality data of ARPA monitoring network in accordance with the ISPRA guidelines to determine the groundwater NBL for As, Fe, Mn and NH4.

  Exploratory analysis was performed trough multivariate statistical analysis in order to identify different groundwater hydrofacies and to reach a better understanding of the conceptual model. A MATLAB script was developed to perform automatically univariate statistical test over single monitoring stations to determine the presence of outliers, trend and the distribution of the data population.

## A.1 Articles

1. Perfido, D., Raciti, M., **Zanotti**, C., Chambers, N., Hannon, L., Keane, M., et al. (2017). Towards sustainable water networks: automated fault detection and diagnosis. ENTREPRENEURSHIP AND SUSTAINABILITY ISSUES, 4(3), 339-350.
2. Stefania, G., **Zanotti**, C., Bonomi, T., Fumagalli, L., & Rotiroti, M. (2018). Determination of trigger levels for groundwater quality in landfills located in historically human-impacted areas. WASTE MANAGEMENT, 75, 400-406.
3. Stefania, G., Rotiroti, M., Fumagalli, M., Simonetto, F., Capodaglio, P., **Zanotti**, C., et al. (2018). Modeling groundwater/surface-water interactions in an Alpine valley (the Aosta Plain, NW Italy): the effect of groundwater abstraction on surface-water resources. HYDROGEOLOGY JOURNAL, 26(1), 147-162.
4. Stefania, G., Rotiroti, M., Fumagalli, M., **Zanotti**, C., & Bonomi, T. (2018). Numerical Modeling of Remediation Scenarios of a Groundwater Cr(VI) Plume in an Alpine Valley Aquifer. GEOSCIENCES, 8(6).
5. Rotiroti, M., Bonomi, T., Sacchi, E., Mcarthur, J., Stefania, G., **Zanotti**, C., et al. (2019). The effects of irrigation on groundwater quality and quantity in a human-modified hydro-system: The Oglio River basin, Po Plain, northern Italy. SCIENCE OF THE TOTAL ENVIRONMENT, 672, 342-356.
6. **Zanotti**, C., Bonomi, T., Nannucci, M., & Rotiroti, M. (2019). Groundwater level forecasting using linear time series modeling: the case study of the thermal aquifer system of Monsummano Terme (central Italy). RENDICONTI ONLINE DELLA SOCIETÀ GEOLOGICA ITALIANA, 47, 153-160.
7. Rotiroti, M., **Zanotti**, C., Fumagalli, L., Taviani, S., Stefania, G., Patelli, M., et al. (2019). Multivariate statistical analysis supporting the hydrochemical characterization of groundwater and surface water: a case study in northern Italy. RENDICONTI ONLINE DELLA SOCIETÀ GEOLOGICA ITALIANA, 47, 90-96.
8. **Zanotti**, C., Rotiroti, M., Fumagalli, L., Stefania, G., Canonaco, F., Stefenelli, G., et al. (2019). Groundwater and surface water quality characterization through positive matrix factorization combined with GIS approach. WATER RESEARCH, 159, 122-134.

9. Stefania, G., Rotiroti, M., Buerge, I., **Zanotti**, C., Nava, V., Leoni, B., et al. (2019). Identification of groundwater pollution sources in a landfill site using artificial sweeteners, multivariate analysis and transport modeling. WASTE MANAGEMENT, 95(15 July 2019), 116-128.
10. **Zanotti**, C., Rotiroti, M., Sterlacchini, S., Cappellini, G., Fumagalli, L., Stefania, G. A., Nannucci, M.S, Leoni, B., Bonomi, T. (2019). Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system. Journal of Hydrology, 124015. In press, Available online.

## A.2 Abstract

11. Rotiroti, M., Bonomi, T., Fumagalli, L., Taviani, S., Stefania, G., **Zanotti**, C., et al. (2017). Using isotopic and hydrochemical data to investigate groundwater recharge and discharge in a highly impacted watershed: the Oglio River, northern Italy. GEOPHYSICAL RESEARCH ABSTRACTS, 19, 8195-8195.
12. Taviani, S., Bonomi, T., Fumagalli, L., Rotiroti, M., Stefania, G., **Zanotti**, C., et al. (2017). Hydrogeological conceptual model of a highly impacted watershed: the case study of Oglio River (N Italy). In G. Ghiglieri, & S. Da Pelo (a cura di), Flowpath 2017, 3rd National Meeting on Hydrogeology, Conference Proceedings (pp. 116-117). Dipartimento di Scienze Chimiche e Geologiche, Università degli Studi di Cagliari.
13. Rotiroti, M., Bonomi, T., Fumagalli, L., **Zanotti**, C., Taviani, S., Stefania, G., et al. (2017). Hydrochemical characterization of groundwater and surface water supported by multivariate statistical analysis: a case study in the Po plain (in Italy). In G. Ghiglieri, & S. Da Pelo (a cura di), Flowpath 2017, 3rd National Meeting on Hydrogeology, Conference Proceedings (pp. 39-40). Dipartimento di Scienze Chimiche e Geologiche, Università degli Studi di Cagliari.
14. **Zanotti**, C., Rotiroti, M., Nannucci, M., & Bonomi, T. (2017). Linear time series modelling for groundwater level forecasting: the case study of the fractured aquifer system of Monsummano terme (central Italy). In Flowpath 2017, 3rd National Meeting on Hydrogeology, Conference Proceedings (pp.80-81). Dipartimento di Scienze Chimiche e Geologiche, Università degli Studi di Cagliari.

15. Rotiroti, M., Bonomi, T., Fumagalli, L., McArthur, J., Sacchi, E., Taviani, S., **Zanotti**, C., et al. (2017). Using Cl/Br ratios and water isotopes to trace aquifer recharge in a highly irrigated area, the Po Plain (N Italy). In K. Posavec, & T. Marković (a cura di), 44th IAH Congress - Book of Abstracts (pp. 235-235). IAH.
16. Nava, V., Patelli, M., **Zanotti**, C., Rotiroti, M., Stefania, G., Soler, V., et al. (2018). Effect of hydrological variations and land-use evolution on chloride trend: a long-term analysis in the Oglio River-Lake Iseo system (Northern Italy). Intervento presentato a: Congresso Società Italiana di Ecologia - SItE, Cagliari, Italy.
17. Stefania, G., **Zanotti**, C., Rotiroti, M., Fumagalli, L., Bigoni, M., Simonetto, F., et al. (2018). Assessment of groundwater quantitative status: implementation on the Aosta Plain aquifer. In Atti del Convegno del 6° Congresso Nazionale AIGA 2018 (pp.6-6).
18. Nava, V., Patelli, M., **Zanotti**, C., Rotiroti, M., Stefania, G., Soler, V., et al. (2018). Long-term chloride quantification in a mixed-land-use watershed (Iseo-Oglio system). In XXXIV Congress International Society of Limnology (SIL), August 19-24, 2018 – Nanjing, China, Book of Abstracts (pp.164-165). International Society of Limnology.
19. Nava, V., Patelli, M., **Zanotti**, C., Rotiroti, M., Stefania, G., Soler, V., et al. (2018). Chloride increase in the deep south-Alpine lake Iseo (Northern Italy): load quantification and source identification with a watershed approach. Intervento presentato a: ELLS-IAGLR - Big Lakes, Small World, Evian-les-bains, France.
20. **Zanotti**, C., Bonomi, T., Fumagalli, L., Rotiroti, M., Stefania, G., Taviani, S., et al. (2019). Groundwater and surface water quality characterization with positive matrix factorization. In L. Alberti, T. Bonomi, & M. Masetti (a cura di), Flowpath 2019, National Meeting on Hydrogeology - Conference Proceedings (pp. 64-66). Ledizioni.
21. Rotiroti, M., Bonomi, T., Fumagalli, M., Sacchi, E., Stefania, G., **Zanotti**, C., et al. (2019). Groundwater/surface-water interactions and their effect on nitrate pollution in the Oglio River basin (N Italy). In J.J. Gómez Hernández, & B. Andreo Navarro (a cura di), Proceedings of IAH2019 (pp. 678-678). Asociación Internacional de Hidrogeólogos – Grupo Español.