ECWS

MDPI

Prediction of Phytoplankton Biomass in Small Rivers of Central Spain by Data Mining Method of Partial Least-Squares Regression

Juan Carlos García Prieto [1,*], Francisco Javier Burguillo Muñoz [2], Manuel García Roig [1], and José Bernardo Proal Najera [3]

[1] Centro de Investigación y Desarrollo Tecnológico del Agua (CIDTA) Universidad de Salamanca; Campus Miguel de Unamuno Facultad de Farmacia s/n Salamanca (Spain); mgr@usal.es
[2] Departament de Química Física; Universidad de Salamanca; Campus Miguel de Unamuno Facultad de Farmacia s/n Salamanca (Spain); burgui@usal.es
[3] Instituto Politécnico Nacional, CIIDIR–Unidad Durango. Sigma 119, Fracc. 20 de Nov. II. 34220. Durango, Dgo., México; joseproal@hotmail.com
* Correspondence: jcgarcia@usal.es; Tel.: +34-923-294-670

Abstract: The Water Framework Directive (WFD, EC, 2000) states that the "good" ecological status of natural water bodies must be based on their chemical, hydromorphological and biological features, especially under drastic conditions of floods or droughts. Phytoplankton is considered a good environmental bioindicator (WFD) and the Climate Change has a strong impact on phytoplankton communities and water quality. The development of robust techniques to predict and control phytoplankton growth is still in progress. The aim of this study is to analyze the impact of the different stressors associated with the change in phytoplanktonic communities in small rivers in the center of the Iberian Peninsula (Southwestern Europe). A statistical study on the identification of the essential limiting variables in the phytoplankton growth and its seasonal variation by Climate Change was carried out. In this study, a new method based on the partial least-squares (PLS) regression technique has been used to predict the concentration of phytoplankton and cyanophytes from 22 variables usually monitored in rivers. The predictive models have shown a good agreement between training and test data sets in rivers and seasons (dry and wet). The phytoplankton in dry periods showed greatest similarities, being these dry periods the most important factor in the phytoplankton proliferation

Keywords: phytoplankton; climate change; prediction; Partial Least Squares Regression

1. Introduction

The Water Framework Directive (WFD, EC, 2000) states that the "good" ecological status of natural water bodies must be based on their chemical, hydromorphological and biological characteristics, compared to the reference conditions [1]. To comply with the protection of surface waters established in the Water Framework Directive, it is necessary to monitor the ecological and chemical status of water quality, especially under drastic conditions of floods or droughts due to the greater epidemiological risk that occur during these periods.

Phytoplankton is considered a good environmental bioindicator since it presents temporary patterns related to environmental changes and, in addition, the processes that act on this community operate on a reduced time scale, so phytoplankton is an important ecological tool to obtain answers in the short term [2,3]. Furthermore, spatio-temporal variability in the structure of phytoplankton communities plays an important role in the structure and function of aquatic ecosystems [4]. Multiple factors affect the phytoplankton population, among these are the main nutrients (nitrogen, carbon and phosphorus) [5], the environmental conditions, the hydrodynamics and hydromorphology of rivers [6,7] and the biotic conditions (competition, predators, etc.) [8].

With regards to environmental and climatic conditions, phytoplankton depends on light intensity and temperature since they affect the speed of photosynthetic processes [9; 10], on the level of the water surface since a low flow rate and a decrease in the level of water in rivers produces an increase of phytoplankton [11]. Other studies have also shown that increasing organic carbon and nutrient inputs from landfills can lead to changes in the competitive dynamics between bacteria and phytoplankton, reducing phytoplankton biomass and increasing bacterial abundance [5]. In this sense, Climate Change affects ecosystems on a planetary scale [12] and is especially important in some regions around the world. Thus, several predictive models have shown that the Mediterranean climate region is particularly sensitive to global warming due to the progressive establishment of a drier and warmer climate [13,14]. The effects of drought on the hydrology of the Mediterranean basins has been studied [15–19] since it is expected that the effects - in terms of frequency and intensity - of the hydrological drought will be more severe due to Climate Change.

The Climate Change has a special effect on unregulated rivers that are temporary or intermittent. Temporary rivers are ecologically unique, supporting important ecosystem processes and functions and being highly relevant in the conservation and protection of the biodiversity. At the same time, they suffer a large number of anthropogenic impacts, including alterations of their flow regime, changes in their bends and channels, nutrients excesses and invasive species [20]. Predictions on Climate Change have indicated that the Mediterranean region will suffer severe deficits in the flow of its rivers, increasing the vulnerability of temporary rivers and of those that are now perennial, which will become temporary [21,22]. The appropiate management of the rivers, maintaining their level and flow in regulated rivers, can improve the quality of the water, especially when they contain phytoplankton species that can harm the human population such as cyanobacteria [23].

The objective of this study is to analyze the impact of the different stressors associated with the change in phytoplanktonic communities in small rivers in the center of the Iberian Peninsula (Southwestern Europe) with the multivariate method of Partial Least Squares (PLS). PLS statistical regression is a recent technique that generalizes and combines features from principal component analysis and multiple regression [24,25] and that can be used to analyze data from environmental effects on biodiversity [26,27] and large-scale influence of climate [28,29]. In the present study, the establishment of statistical models, suitable for predicting concentration of phytoplankton and cyanophytes from 22 variables usually monitored in rivers, has been carried out. Furthermore, the influence of phytoplankton and cyanobacteria concentration with respect to other environmental and morphological variables in the different sampling points and seasonal periods, has been established. A better knowledge of the limiting factors in the growth of phytoplankton will allow watershed managers to improve the quality of the discharge sites and prevent risks to the population.

2. Materials and Methods

2.1. Study Area

The study area for the determination of superficial water quality is located in the province of Salamanca (Western Spain). This province covers an area of 12 340 km^2 and forms the South-Western part of the River Duero basin, which is the most important aquifer system of the Iberian Peninsula. The climate of the region is continental, with considerable seasonal fluctuations in temperature (the difference in mean temperature between the hottest and coldest days is almost 20 °C) and low humidity. Precipitation is low (mean annual rainfall 380 mm), highly irregular and usually absent in July and August, and, hence, during the dry season the hydric balance is clearly negative. This Salamanca province has 3 river basins (Figure 1), two belonging to the Duero river, (Tormes and Águeda river basins) and one river basin belonging to Tajo river (Alagón river basin). The Tormes river basin is not contemplated because it has been previously studied in depth by the authors [18].

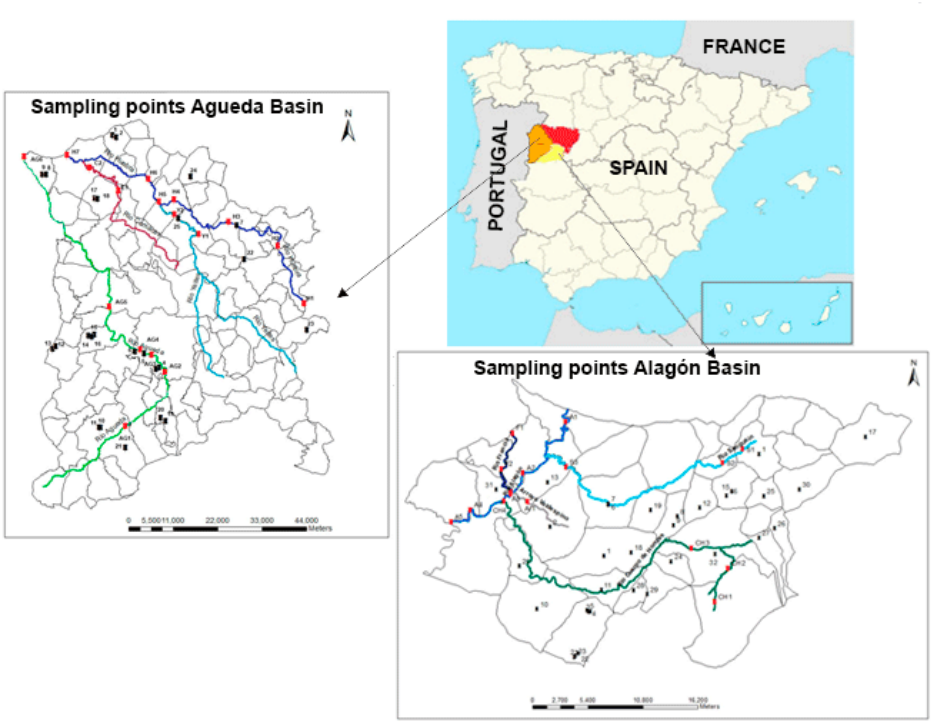

Figure 1. Geographic locations of the sampling sites on Agueda and Alagón rivers where data were collected.

2.2. Sampling and Analysis

The 22 parameters were measured at 33 sampling points (Figure 1: red points). They were selected to evaluate the evolution of the quality of water of the Agueda and Huebra rivers (Agueda river basin) and Alagón river (Alagón river basin) upstream and downstream of municipal wastewater discharges (Figure 1: black points) to consider the influence of these discharges on water quality. The present study has been carried out during the years 2015 and 2017. Furthermore, within the years studied, 2 seasonal periods have been investigated. May to September seasonal period is considered as summer (summer 2015 and 2017) and November to March seasonal period as winter (Winter 2017). On the other hand, the first study period corresponds to the 2014-2015 hydrological year, been considered as a wet hydrological year. The second period corresponding to the year 2017 (hydrological years 2016-2017 and 2017-2018) registered a rainfall much lower than normal, having been considered as very dry period. This covered an extreme drought occurring from mid-July 2016 until mid-October 2017.

The analyses parameters were: total solids, ammonia, nitrite, nitrate, total phosphorus, sulfate, chloride, fluoride, calcium, magnesium, chemical oxygen demand, biochemical oxygen demand, total organic carbon, colour and total and fecal coliforms in the water samples. This parameters were determined using official or recommended methods of analysis [29,30]. The in situ measurements were: pH, temperature, conductivity, turbidity, and dissolved oxygen. Algal class analysis (Cyanophyta, Cryptophyta, Chlorophyta, Bacillariophyta and Dinophyta) was carried out with the fluoroprobe, a submergible spectrofluorometer (bbe FluoroProbe) [31].

2.3. PLS Regression Method

The prediction models were set up using of the PLS option of SIMFIT statistical open source package [32]. PLS regression is particularly useful to predict a new set of dependent variables (response) from a large set of independent variables (predictors). Prediction models are achieved by extracting from the predictors and response variables a new set of orthogonal factors called latent variables, which capture the best predictive power. PLS regression searches for a set of components

performing a simultaneous decomposition of predictors and response variables with the constraint that these components explain as much as possible the covariance between predictors and responses

3. Results

Two river basins at two different seasonal periods (dry and wet) have been studied. As an example, the development of the predictive model for the dry winter period in the Águeda River is presented.

The PLS technique considers two types of matrices of variables, on the one hand the matrix of predictive variables (X) that will be composed, for each of the rivers in each of the stations studied, by the values of the 22 variables measured. On the other hand, the matrix of response variables (Y) encompasses the two variables to be predicted, which are phytoplankton, measured chlorophyll-a, and cyanobacteria, measured as phycocyanin pigment. Figure 2 shows the cumulative variance of the latent factors, for both the X and Y variables in the Agueda river (dry winter seasonal period). As can be seen, a plateau is reached where the gain in capturing the variability is very small. Based on the fact that this capture of variability is considered acceptable, can be admitted for calibration purposes that 7 factors are sufficient for the model (97% capture of variability in X and 92.94% in Y for phytoplankton and 97% in X and 89% in Y capture variability for cyanobacteria).

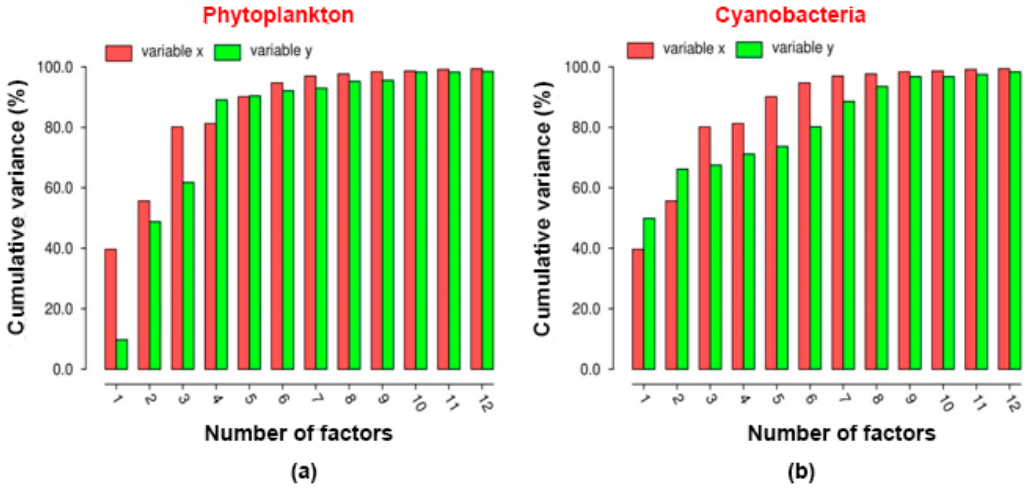

Figure 2. Cumulative variance captured against the number of PLS factors, for both phytoplankton (**a**) and cyanobacteria (**b**).

To quantify the importance of each of the variables X in the prediction model, the scores of statistics VIP ("Variable Influence on Projection" [32]) was used. The VIP-scores for the 22 variables X put into play, for prediction model built with 7 factors, are shown in Figure 3. Important predictors were identified in the modelling of phytoplankton and cyanobacteria concentration by considering the variables with VIP-scores higher than one. It should be highlighted as better predictors for both are temperature, ammonium and fecal coliforms.

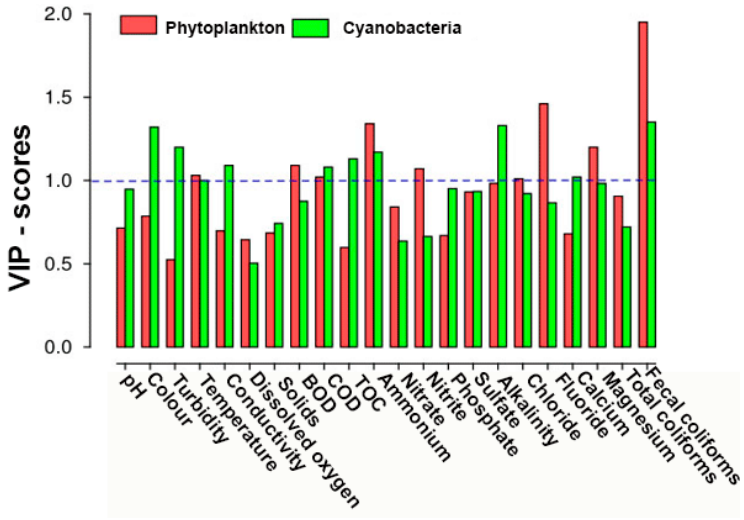

Figure 3. VIP-scores of the PLS prediction model for phytoplankton and cyanobacteria concentration in the dry winter seasonal period of Águeda river.

PLS, as well as its interpretation, can be expressed in the form of a multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{21} X_{21} + \beta_{22} X_{22}$$

Being β_i (from i = 1 to 22) the coefficients of the predictive variables in the constructed model, and β_0 the independent term. Finally, the following equations are obtained

Phytoplankton = 61,3798 - 1,0141 pH + 0,1233 Colour + 0,0550 Turbidity - 1,7876 Temperature + 0,0007 Conductivity - 0,0731 Dissolved oxygen + 0,0009 Solids - 0,3679 BOD - 0,1532 COD - 0,2751 TOC + 4,7291 Ammonium + 0,7761 Nitrate -74,8815 Nitrite -13,8925 Phosphate + 0,0389 Sulfate - 0,0226 Alkalinity + 0,1323 Chloride - 13,2087 Fluoride - 0,0312 Calcium + 0,5091 Magnesium - 0,0003 Total Coliforms + 0,0040 Fecal Coliforms.

Cyanobacteria = 41,6068 - 1,4866 pH + 0,1075 Colour - 0,0438 Turbidity - 1,6663 Temperature - 0,0038 Conductivity + 0,5793 Dissolved oxygen + 0,0050 Solids - 0,1817 BOD - 0,0628 COD - 0,2621 TOC + 19,6949 Ammonium - 0,4312 Nitrate - 46,0138 Nitrite - 11,3262 Phosphate + 0,0236 Sulfate + 0,0124 Alkalinity + 0,0906 Chloride - 5,2369 Fluoride - 0,0910 Calcium + 0,2238 Magnesium - 0,00001 Total Coliforms+ 0,0027 Fecal Coliforms.

PLS methodology consists of two differentiated parts; calibration with a training-set data and validation with a test-set data. The experimental data for the Ageda river example were divided ¾ for a training-set data and ¼ for the test-set data. The process of calibration (Figure 4) and validation (Table 1) of the model is exposed.

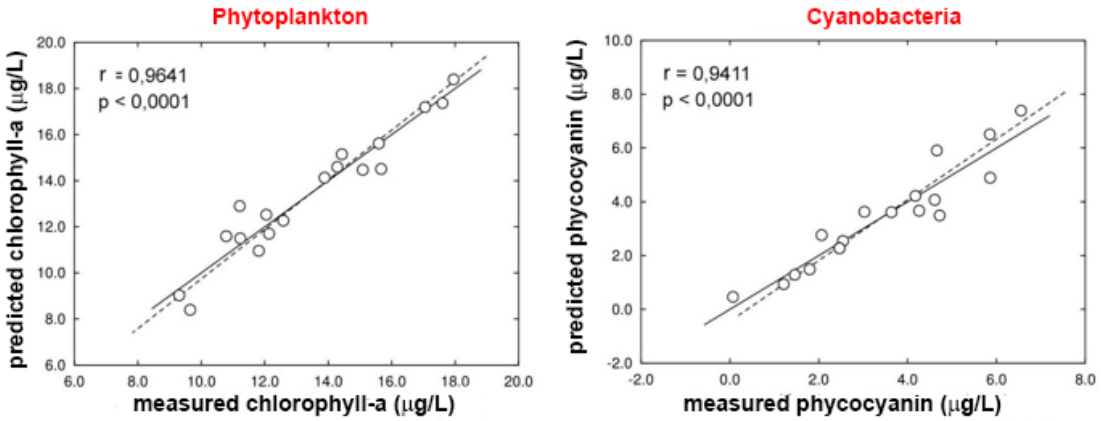

Figure 4. Correlations of the values predicted by the PLS model with the experimental values.

In training procedure 7 PLS factors were selected as are optimum and the agreement between the measured and the predicting values for the model are shown in Figure 4 where it can be seen a good correlation with the training data. Nevertheless, the above good agreement with the training set data is not the better approach for the goodness of the model. Therefore, a test-set with a new experimental data were used to validate the model. The prediction rates for the sampling points in the Agueda river example are presented in Table 1.

Table 1. Validation of PLS phytoplankton and cyanobacteria concentrations models.

	Phytoplankton			*Cyanobacteria*		
Samplig Points	**Y real (μg/L)**	**Y predicted (μg/L)**	**Relative Error (%)**	**Y real (μg/L)**	**Y predicted (μg/L)**	**Relative Error (%)**
Irueña	13.57	13.47	0.69	1.73	2.36	36.84
Sanjuanejo	18.03	15.90	11.81	6.73	6.10	9.38
C. Rodrigo	16.31	12.03	26.27	5.79	2.95	48.99
Ivanrey	11.48	10.40	9.40	2.31	2.73	18.19
Siega Verde	14.83	12.78	13.80	2.14	3.39	58.08
Fregeneda	15.5	11.49	25.86	3.76	1.11	70.41
Average			**14.64**			**40.31**

As shown on Table 1, the prediction error percentages have been better for phytoplankton (15%) than for cyanobacteria (40%), which indicates a good fit of the PLS prediction model for phytoplankton.

4. Discussion

Following the same methodology, in order to carry out some comparisons between the rivers, all the studies were carried out using the same PLS statistical procedure with 7 factors for the different rivers in the different seasonal periods analyzed. The results of the comparison are shown in the conclusions.

5. Conclusions

A new methodology, based on the multivariate regression technique PLS, has been proposed in this work, which allows, based on 22 variables usually monitored in rivers, to predict the concentration of phytoplankton and cyanophytes. The predictive models generated have presented a goodness of fit tested successfully using training data series. In turn, these models have performed well for the prediction of phytoplankton and cyanobacterial concentrations from new validation data series, although prediction error rates have been better for phytoplankton (10-25%) than for cyanobacteria (40-60%).

Predictive models are formulated by equations of the linear multiple regression type where the coefficients indicate the participation of each of the variables in the model. In this sense, the determined coefficients have varied from one river to another and between seasons, what was expected. However, a certain similarity of the coefficients for dry summer periods (droughts) has been observed. In these transition periods, their features are most important in the prediction, since they exhibit favourable conditions for the proliferation of the phytoplankton community

References

1. Feio, M.J., Aguiar, F.C., Almeida, S.F.P., Ferreira, J., Ferreira, M.T., Elias, C., Serra, S.R.S., Buffagni, A., Cambra, J., Chauvin, C., Delmas, F., Dörflinger, G., Erba, S., Flor, N., Ferréol, M., Germ, M., Mancini, L., Manolaki, P., Marcheggiani, S., Minciardi, M.R., Munné, A., Papastergiadou, E., Prat, N., Puccinelli, C., Rosebery, J., Sabater, S., Ciadamidaro, S., Tornés, E., Tziortzis, I., Urbanič, G., Vieira, C.. Least disturbed

condition for European Mediterranean rivers. Sci. Total Environ., 2014, 476–477, 745–756. https://doi.org/10.1016/j.scitotenv.2013.05.056
2. Reynolds, C.S. Ecological pattern and ecosystem theory. Ecological Modelling, 2002, 158(3), 181-200. https://doi.org/10.1016/S0304-3800(02)00230-2
3. Wetzel, R. G. Limnology. Lake and River Ecosystems. Third Edition. Academic Press, USA. 2001 p.1006
4. Brett MT, Goldman CR. A meta-analysis of the freshwater trophic cascade. PNAS USA 1996, 93: 7723-7726. https://doi.org/10.1073/pnas.93.15.7723
5. Carney R. L.; Seymour J. R. ; Westhorpe D. and Mitrovic S. M. Lotic bacterioplankton and phytoplankton community changes under dissolved organic-carbon amendment: evidence for competition for nutrients Mar Freshwater Res .2016, 67(9) 1362-1373 https://doi.org/10.1071/MF15372
6. Hallegraeff GM. A review of harmful algae blooms and their apparent global increase. Phycologia 1993, 32: 79-99. https://doi.org/10.2216/i0031-8884-32-2-79.1.
7. Humborg C, Ittekkot V, Cociasu A, Bodungen BV. Effect of Danube river dam on black sea biogeochemistry and ecosystem structure. Nature 1997, 386: 385-388. https://doi.org/10.1038/386385a0
8. Hutchinson GE. The paradox of the plankton. The American Naturalist, 1961 95: 137-145 https://doi.org/10.1086/282171
9. Sand-Jensen K, Borum J. Interactions among phytoplankton, periphyton, and macrophytes in temperate freshwaters and estuaries. Aquatic Botany 1991, 41,137–75 https://doi.org/10.1016/0304-3770(91)90042-4
10. Béchet Q., Shilton A., Guieysse B. Modeling the effects of light and temperature on algae growth: State of the art and critical assessment for productivity prediction during outdoor cultivation. Biotech Adv 2013, 31, (8), 1648-1663 https://doi.org/10.1016/j.biotechadv.2013.08.014
11. Vis C., Hudon C. Carignan R. and Gagnon P. Spatial Analysis of Production by Macrophytes, Phytoplankton and Epiphyton in a Large River System under Different Water-Level Conditions. Ecosystems 2007, 10, 293–310 https://doi.org/10.1007/s10021-007-9021-3
12. Briner, S., Elkin, C., Huber, R. Evaluating the relative impact of climate and economic changes on forest and agricultural ecosystem services in mountain regions. J Environ Manag. 2013, 129, 414–422. https://doi.org/10.1016/j.jenvman.2013.07.018
13. Sánchez, E., Gallardo, C.; Gaertner, M. A.; Arribas A. & Castro, M. Future climate extreme events in the Mediterranean simulated by a regional climate model: a first approach. Glob. Planet. Chang. 2004, 44: 180–183. https://doi.org/10.1016/j.gloplacha.2004.06.010
14. Ceballos-Barbancho. A.; Morán- Tejeda, E.; Luengo-Ugidos, M.A.; Llorente-Pinto, J.M; Water resources and environmental change in a Mediterranean environment: The south-west sector of the Duero river basin (Spain). J. Hydrol., 2008 351, (1–2), 30, 126-138 https://doi.org/10.1016/j.jhydrol.2007.12.004
15. Mimikou M.A., Baltas E., Varanou E., and Pantazis K. Regional Impacts of Climate Change on Water Resources Quantity and Quality Indicators J. Hydrol 2000. 234, 95–109 https://doi.org/10.1016/S0022-1694(00)00244-4
16. Cidu R. and Biddau R. Transport of trace elements under different seasonal conditions: Effects on the quality of river water in a Mediterranean area, Appl. Geochem, 2007, 22(12), 2777-2794 https://doi.org/10.1016/j.apgeochem.2007.06.017
17. Giorgi, F., Lionello P. Climate change projections for the Mediterranean region Glob. Planet. Chang.,2008, 63, 90-104 https://doi.org/10.1016/j.gloplacha.2007.09.005
18. García-Prieto, J.C.; Cachaza, J.M., Pérez-Galende, P., García Roig, M. Impact of drought on the ecological and chemical status of surface water and on the content of arsenic and fluoride pollutants of groundwater in the province of Salamanca (Western Spain) Chem. Ecol., 2012, 28(6),1-16 https://doi.org/10.1080/02757540.2012.686608
19. Barceló, D. and Sabater, S. Water quality and assesment under scarcity: Prospects and challenges in Mediterranean watersheds. J. Hydrol 2010, 383, 1-4. https://doi.org/10.1016/j.jhydrol.2010.01.010
20. Han, H., Allan, J.D., Scavia, D. Influence of climate and human activities on the relationship between watershed nitrogen input and river export. Environ. Sci. Technol., 2009, 43 (6), https://pubs.acs.org/doi/10.1021/es801985x
21. Karaouzas, I., Smeti, E., Vourka, A., Vardakas, L., Mentzafou, A., Tornés, E., Sabater, S., Muñoz, I., Skoulikidis, N.T., Kalogianni, E. Assessing the ecological effects of water stress and pollution in a temporary river - Implications for water management. Environ. Sci. Technol., 2018, 618, 1591-1604 https://doi.org/10.1016/j.scitotenv.2017.09.323

22. Skoulikidis, T.N., Sabater, S., Datry, T., Morais, M., Buffagni, A., Dörflinger, G., Zogaris, S., Sánchez-Montoya, M.M., Bonada, N., Kalogianni, E., Rosado, J., Vardakas, L., De Girolamo, A.M., Tockner, K. Non-perennial Mediterranean rivers in Europe: status, pressures, and challenges for research and management. Sci. Total Environ. 2017, 577 (15), 1–18 https://doi.org/10.1016/j.scitotenv.2016.10.147
23. Webster IT, Sherman BS, Bormans M, Jones G. Management strategies for cyanobacterial blooms in an impounded lowland river. Regul. Rivers: Res. Manage., 2000 16: 513-525. https://doi.org/10.1002/1099-1646(200009/10)16:5<513::AID-RRR601>3.0.CO;2-B
24. Abdi, H.. Partial least square regression (PLS regression). Salkind, N. J. (ed.), Encyclopedia of measurement and statistics. Sage. 2007
25. Wold, H.. Soft modelling by latent variables; the nonlinear iterative partial least squares approach. Gani, J. (ed.), Perspectives in probability and statistics. Papers in honour of M. S. Barlett. Academic Press, 1975 pp. 117–142.
26. Palomino, D. and Carrascal, L. M. Habitat associations of a raptor community in a mosaic landscape of central Spain under urban development. Landscape Urban Plan. 2007, 83: 268–274. https://doi.org/10.1016/j.landurbplan.2007.04.011
27. Potapova, M. G. Charles, F. D., Ponader, K.C., Winter, D. M. Quantifying species indicator values for trophic diatom indices: a comparison of approaches. – Hydrobiologia, 2004, 517: 25–41. https://doi.org/10.1023/B:HYDR.0000027335.73651.ea.
28. Finsinger, W. Heiri, O., Valsecchi, V., Tinner W. and Lotter, A.F.. Modern pollen assemblages as climate indicators in southern Europe. – Global Ecol. Biogeogr. 2007 16: 567–582. https://doi.org/10.1111/j.1466-8238.2007.00313.x.
29. AENOR, Calidad del Agua, Asociación Española de Normalización y Certificación, Madrid, 2005.
30. APHA-AWWA-WPCF, Métodos normalizados para el análisis de aguas potables y residuales, 17th ed, 1992.
31. Catherine, A., Escoffier, N., Belhocine A., Nasri A.B., Hamlaoui S., Yepremian C., Bernard C., Troussellier M. On the use of the FluoroProbe, a phytoplankton quantification method based on fluorescence excitation spectra for large-scale surveys of lakes and reservoirs, 2012, Water Res., 46, 1771–1784. https://doi.org/10.1016/j.watres.2011.12.056
32. Wold, S. PLS for Multivariate Linear Modeling QSAR: Chemometric Methods in Molecular Design. In Methods and Principles in Medicinal Chemistry, van de Waterbeemd, H. (ed), : Verlag-Chemie. 1994.