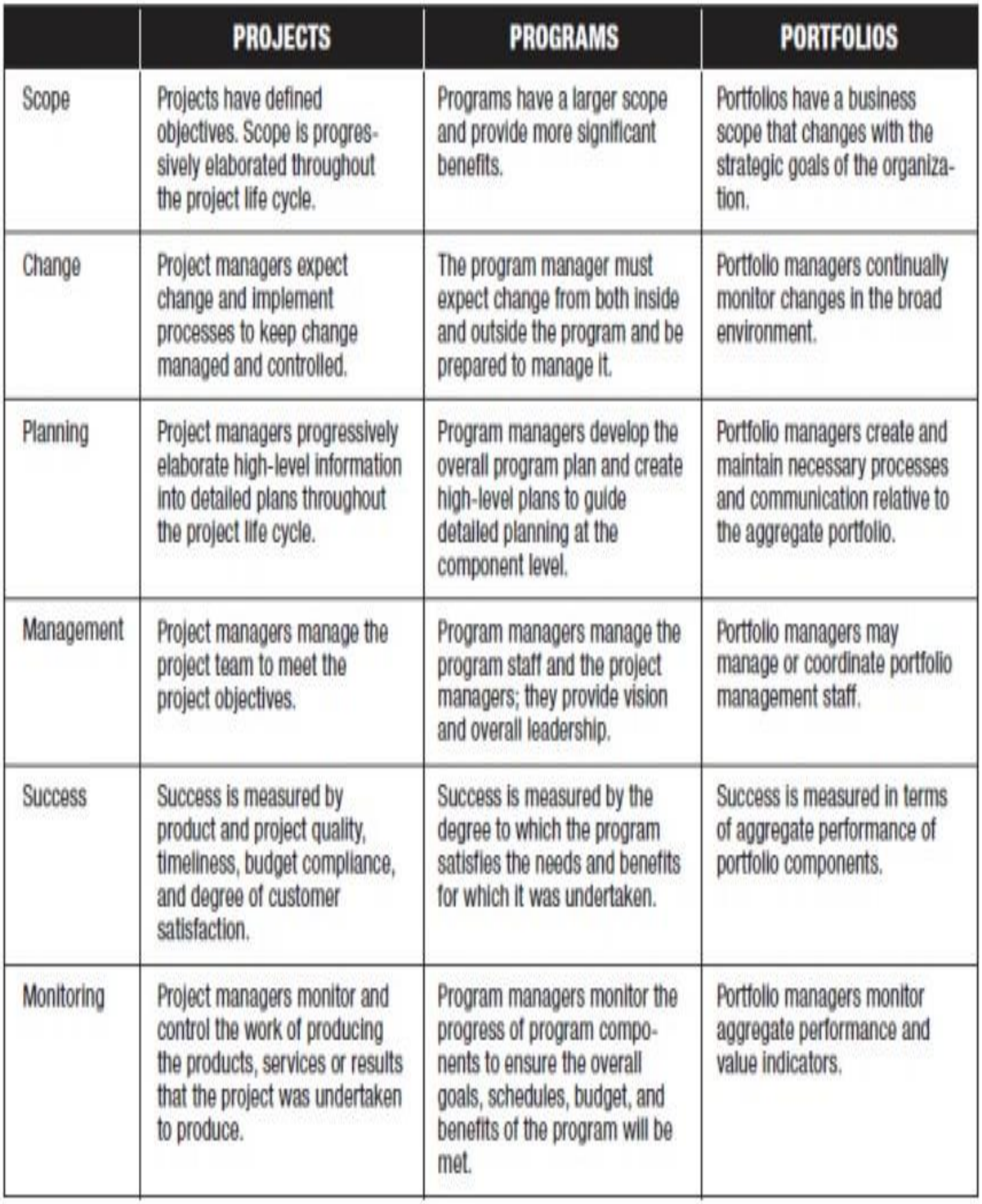
LITERATURE SURVEY

Introduction

PROBLEM STATEMENT:

Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot. The alarming consequences of poor water quality necessitate an alternative method, which is quicker and inexpensive. With this

motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI. The proposed methodology employs four input parameters, namely, temperature, turbidity, pH and total dissolved solids. Of all the employed algorithms, gradient boosting, with a learning rate of 0.1 and polynomial regression, with a degree of 2, predict the WQI most efficiently, having a mean absolute error (MAE) of 1.9642 and 2.7273, respectively. Whereas multi-layer perceptron (MLP), with a configuration of (3, 7), classifies the WQC most efficiently, with an accuracy of 0.8507. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real time water quality detection systems.



	PROJECTS	PROGRAMS	PORTFOLIOS
Scope	Projects have defined objectives. Scope is progressively elaborated throughout the project life cycle.	Programs have a larger scope and provide more significant benefits.	Portfolios have a business scope that changes with the strategic goals of the organization.
Change	Project managers expect change and implement processes to keep change managed and controlled.	The program manager must expect change from both inside and outside the program and be prepared to manage it.	Portfolio managers continually monitor changes in the broad environment.
Planning	Project managers progressively elaborate high-level information into detailed plans throughout the project life cycle.	Program managers develop the overall program plan and create high-level plans to guide detailed planning at the component level.	Portfolio managers create and maintain necessary processes and communication relative to the aggregate portfolio.
Management	Project managers manage the project team to meet the project objectives.	Program managers manage the program staff and the project managers; they provide vision and overall leadership.	Portfolio managers may manage or coordinate portfolio management staff.
Success	Success is measured by product and project quality, timeliness, budget compliance, and degree of customer satisfaction.	Success is measured by the degree to which the program satisfies the needs and benefits for which it was undertaken.	Success is measured in terms of aggregate performance of portfolio components.
Monitoring	Project managers monitor and control the work of producing the products, services or results that the project was undertaken to produce.	Program managers monitor the progress of program components to ensure the overall goals, schedules, budget, and benefits of the program will be met.	Portfolio managers monitor aggregate performance and value indicators.

Environmental Informatics:

Generally, this research may be associated with the new and currently rapidly growing field of environmental informatics. Being one of the directions of the development of data sciences, environmental informatics covers researches that work with data about the state of Earth’s biosphere (and associated spheres) and those processes affecting it. Thus, being interested in reviewing and analysing more projects and articles in this field, one should consider searching for information primarily in this particular area. (Frew and Dozier 2012)

Machine Learning:

Since times of Bayes’ theorem, data mining has greatly developed (especially since the beginning of the computer age) and Machine Learning separated from it as an independent scientific field. There are two the most common definitions of this term. First is provided by 3 Arthur Samuel In 1959, who described it as a "Field of study that gives computers the ability to learn without being explicitly programmed" (Simon 2013). And the more formal definition by Tom M. Mitchell:"A computer program, said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E" (Mitchell 1998). One can define also the following main steps of the analysis using machine learning models:

1. Data Understanding – before defining the possible approaches to work with data, it is necessary to analyse the raw data itself first. What kind of measurements are included, is there any missing data (and in case of natural sciences research, usually there is plenty), which kind of models it is possible to apply to the data and defining the initial goal of the research

2. Data Preparation – merging data, imputing missing values or excluding variables with too many missing values, sorting data, etc.

3. Model Training – actually training the models and analyzing data

4. Results Evaluation – an important stage of the results understanding, which makes possible adjustment of the models and correction of the initial research plan (Chapman, et al. 2000) Additionaly, it is worth defining and explaining the main types of models one can apply:

1. Supervised learning - these are methods where a given set of independent variables are to be matched to one or more dependent variables. During this kind of analysis, model is given a "labled data", where it can find the real values of the parameter it is working with for some certain measurement and values of other parameters for the same measurement, thus it can fit a function. These can be regression tasks (working with

continuous values) and classification tasks (working with class labeled data)

2. Unsupervised learning - in contrast, with unsupervised methods there is no prior “correct” data and the purpose of this kind of analysis is to search for the underlying patterns in the data

3. Optimization - techniques for finding the optimal set of parameters which minimize a pre-defined cost function.

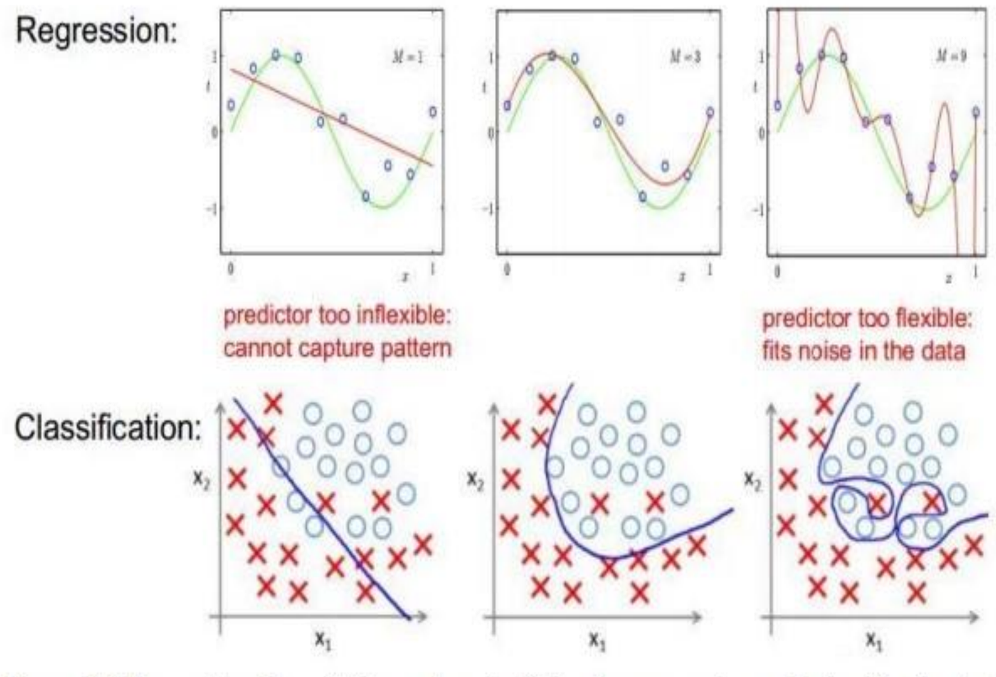


Figure 1.2 Examples of *overfitting* and *underfitting* for *regression* and *classification* tasks

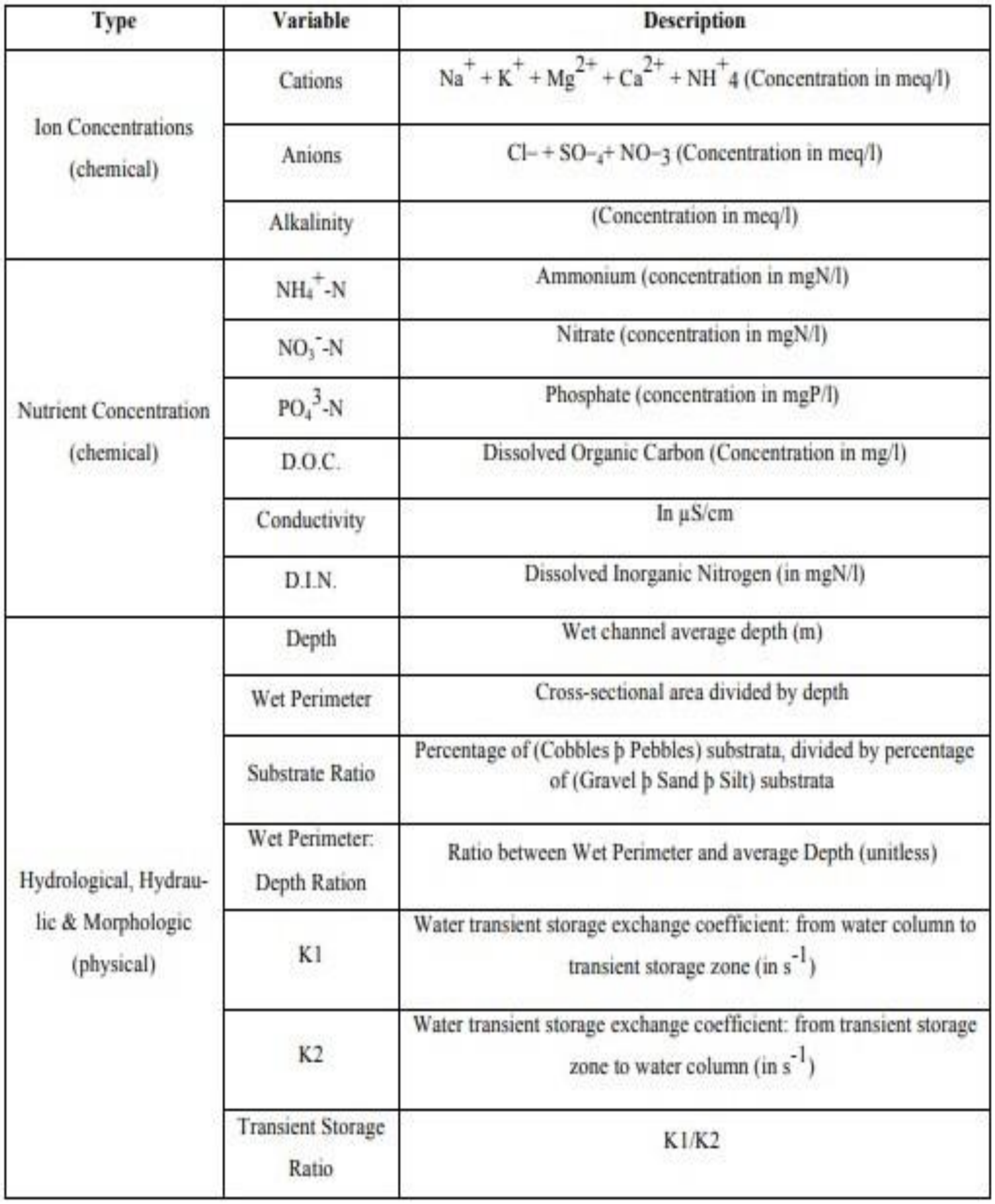
MATERIALS AND METHODS :

This chapter gives information about the data, followed by the explanation of the algorithms and overview of the software used during the analysis.

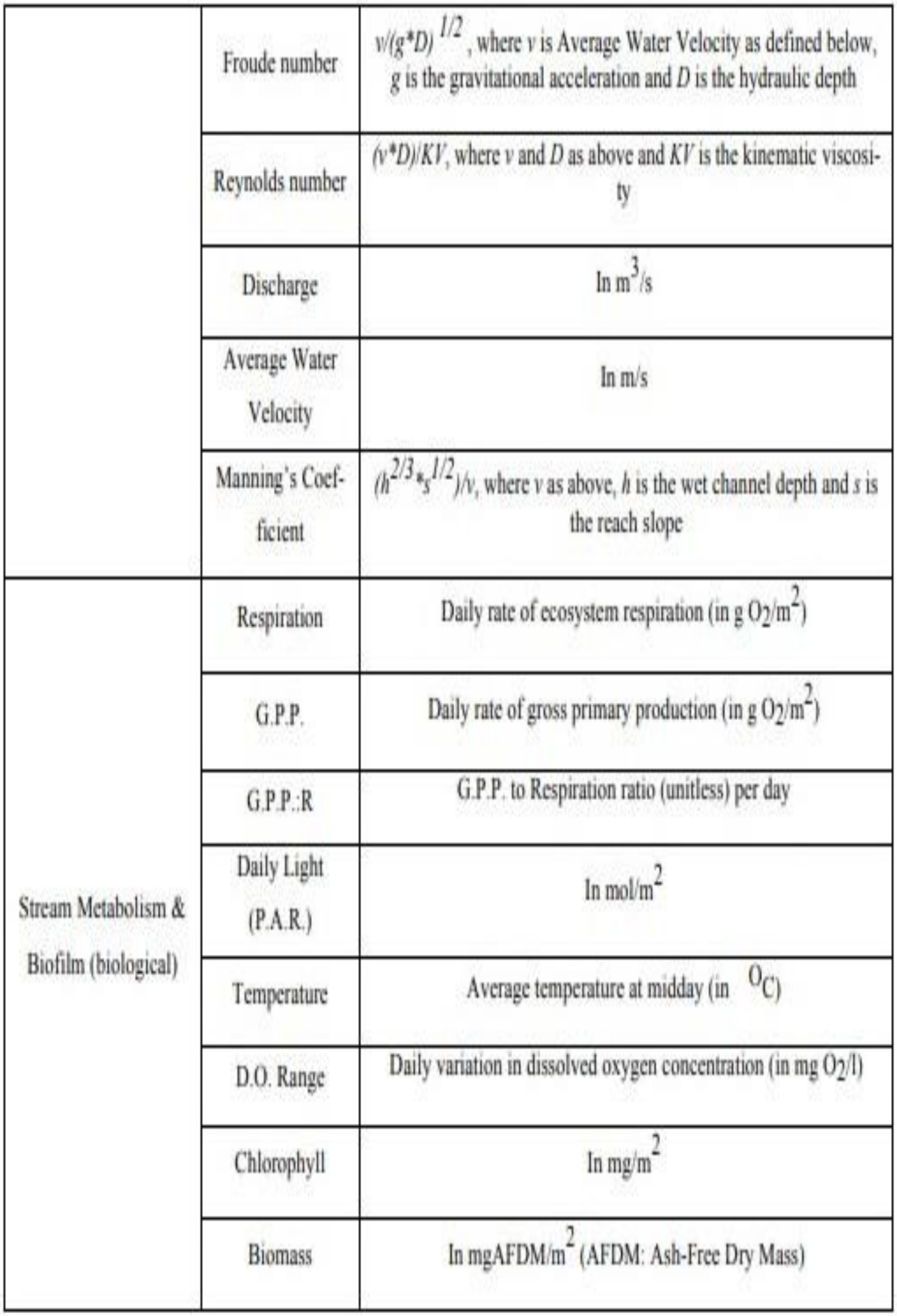
Data

The data used for this research was generated during European STREAMES (Stream reach Management, an Expert System) project, which is an international enterprise for the development of a knowledge-based environmental decision support system to assist water managers with their decision-making tasks. The core of the project itself involved the evaluation of the effect of substantial nutrient loads on the overall water quality and ecological status of stream ecosystems. Empirical data for the knowledge base come from several streams located throughout Europe and Israel, with emphasis on streams from the Mediterranean region. These data comprise several types of variables, including physical, chemical and biological parameters.

Table 2.2 List of the 29 variables selected for the study, grouped by their topology (Vellido, et al. 2007)



Type	Variable	Description
Ion Concentrations (chemical)	Cations	$Na^{+} + K^{+} + Mg^{2+} + Ca^{2+} + NH^{+}4$ (Concentration in meq/l)
	Anions	$Cl- + SO-_{4} + NO-_{3}$ (Concentration in meq/l)
	Alkalinity	(Concentration in meq/l)
Nutrient Concentration (chemical)	NH_{4}^{+}-N	Ammonium (concentration in mgN/l)
	NO_{3}^{-}-N	Nitrate (concentration in mgN/l)
	PO_{4}^{3}-N	Phosphate (concentration in mgP/l)
	D.O.C.	Dissolved Organic Carbon (Concentration in mg/l)
	Conductivity	In µS/cm
	D.I.N.	Dissolved Inorganic Nitrogen (in mgN/l)
Hydrological, Hydraulic & Morphologic (physical)	Depth	Wet channel average depth (m)
	Wet Perimeter	Cross-sectional area divided by depth
	Substrate Ratio	Percentage of (Cobbles þ Pebbles) substrata, divided by percentage of (Gravel þ Sand þ Silt) substrata
	Wet Perimeter: Depth Ration	Ratio between Wet Perimeter and average Depth (unitless)
	K1	Water transient storage exchange coefficient: from water column to transient storage zone (in s^{-1})
	K2	Water transient storage exchange coefficient: from transient storage zone to water column (in s^{-1})
	Transient Storage Ratio	K1/K2



	Froude number	$v/(g^*D)^{1/2}$, where v is Average Water Velocity as defined below, g is the gravitational acceleration and D is the hydraulic depth
	Reynolds number	$(v^*D)/KV$, where v and D as above and KV is the kinematic viscosity
	Discharge	In m^3/s
	Average Water Velocity	In m/s
	Manning's Coefficient	$(h^{2/3}*s^{1/2})/v$, where v as above, h is the wet channel depth and s is the reach slope
Stream Metabolism & Biofilm (biological)	Respiration	Daily rate of ecosystem respiration (in g O_2/m^2)
	G.P.P.	Daily rate of gross primary production (in g O_2/m^2)
	G.P.P.:R	G.P.P. to Respiration ratio (unitless) per day
	Daily Light (P.A.R.)	In mol/m^2
	Temperature	Average temperature at midday (in ^{O}C)
	D.O. Range	Daily variation in dissolved oxygen concentration (in mg O_2/l)
	Chlorophyll	In mg/m^2
	Biomass	In $mgAFDM/m^2$ (AFDM: Ash-Free Dry Mass)

Models and Software:

There is a huge variety of machine learning algorithms and tools existing nowadays. In the following subchapters, the following algorithms used in this research are covered: support vector machines, random forests, artificial neural networks (used for classification, regression, variable importance tasks), k-nearest neighbours (used for data imputation) and k-means clustering (used for unsupervised classification