

# Groundwater Modeling with Machine Learning Techniques: Ljubljana polje Aquifer <sup>†</sup>

Klemen Kenda <sup>1,2,\*</sup>, Matej Čerin <sup>1</sup>, Mark Bogataj <sup>1</sup>, Matej Senožetnik <sup>1,2</sup>, Kristina Klemen <sup>3</sup>, Petra Pergar <sup>3</sup>, Chrysi Laspidou <sup>4</sup> and Dunja Mladenčić <sup>1,2</sup>

<sup>1</sup> Artificial Intelligence Laboratory, Jožef Stefan Institute, Ljubljana 1000, Slovenia; matej.cerin@ijs.si (M.Č.); mark.bogataj18@gmail.com (M.B.); matej.senozetnik@ijs.si (M.S.); dunja.mladenic@ijs.si (D.M.)

<sup>2</sup> Jožef Stefan International Postgraduate School, Ljubljana 1000, Slovenia

<sup>3</sup> Ljubljanski urbanistični zavod, d.d., Ljubljana 1000, Slovenia; kristina.klemen@luz.si (K.K.); petra.pergar@luz.si (P.P.)

<sup>4</sup> Civil Engineering Department, University of Thessaly, Pedion Areos, 38334 Volos, Greece; laspidou@uth.gr

\* Correspondence: klemen.kenda@ijs.si; Tel.: +386-31-420-418

† Presented at the 3rd EWaS International Conference on “Insights on the Water-Energy-Food Nexus”, Lefkada Island, Greece, 27–30 June 2018.

Published: 3 August 2018

**Abstract:** In this study a thorough analysis is conducted concerning the prediction of groundwater levels of Ljubljana polje aquifer. Machine learning methodologies are implemented using strongly correlated physical parameters as input variables. The results show that data-driven modelling approaches can perform sufficiently well in predicting groundwater level changes. Different evaluation metrics confirm and highlight the capability of these models to catch the trend of groundwater level fluctuations. Despite the overall adequate performance, further investigation is needed towards improving their accuracy in order to be comprised in decision making processes.

**Keywords:** groundwater levels; data-driven modeling; machine learning; data preprocessing

## 1. Introduction

To address challenges associated with climate resilience and sustainability principles, the importance of urban groundwater must be integrated into urban planning and design. Groundwater systems are dynamic and adjust continually to short-term and long-term changes in climate, groundwater withdrawal, and land use. Water level measurements from observation wells are the principal source of information about the hydrologic stresses acting on aquifers and how these stresses affect ground-water recharge, storage, and discharge. In this research we focus on *Ljubljana polje* aquifer.

Traditionally groundwater levels are modeled with process-based models, which rely on the profound knowledge of the observed system dynamics. They require many additional spatial data on geological and hydrological properties of the aquifer. On the other hand, in data-driven modeling with machine-learning techniques our model is based solely on the data and some domain-specific knowledge is incorporated in to the system via appropriate data transformation (within engineering of new attributes). The goal in such a scenario would be to predict groundwater levels based on temporal data inputs (historic groundwater and surface water level data, weather data and forecasts, land-use, groundwater withdrawal and other anthropogenic data) and outputs (groundwater level). The model captures underlying processes based on the data without additional expert user input. In our work we present the whole data-mining pipeline, including exploratory data analysis, data pre-processing and modeling, where we explore accuracy and other benefits of a variety of modeling techniques on the same dataset (from interpretable modeling techniques such as

multivariate linear regression, linear SVM and decision trees to black-box models such as gradient boosted trees, random forests and artificial neural networks), which has rarely been reported in scientific literature for water-related scenarios.

Authors of [1] claimed already a decade ago that data-driven modeling has overcome the initial stage and the main objectives shifted from method development and testing to the construction of useful architectures and applications of data-driven modeling for decision makers, according to the availability of the data. In reality, however, machine learning is still seeking its way into the practice and many studies have been published recently, which are researching the usability of different machine learning algorithms. Time-series techniques (ARX, ARMAX, ARMA, ARIMA) and artificial neural network models have been tested in [2]. Random forests and maximum entropy have been tested in [3]. Sahoo et al. [4] focused on multilayer perceptron networks. With  $R^2$  scores higher than 0.8 they claim that data-driven approach can be used as an alternative to process modeling techniques.

The *Ljubljana polje* aquifer has been studied extensively, mainly with process models [4–7]. The aquifer recharges mainly through infiltration of river water, infiltration of precipitation and in smaller fraction through lateral underground flow [8]. The latter is still not studied and understood well [9]. Therefore we can expect that appropriate data to model groundwater will be originating geographically within the aquifer (surface water levels, precipitation, etc.).

Machine learning models have not been studied in the *Ljubljana polje* aquifer, however some preliminary work has been reported on automatic data acquisition [10]. Markov chains have been used to model heterogeneity of the aquifer [11]. Machine learning models have been used to study geological-geomechanical properties of Ljubljana area [12].

## 2. Materials and Methods

### 2.1. Data and Data Acquisition

For data acquisition, we used Internet of Things (IoT) middleware which consists of four main components, named Retrievers, Collector, API Management and Watchdog. The infrastructure is able to retrieve data from the heterogeneous sources, transform it into a unified format, store it and finally expose it via standardized web services, so it is available to users for further processing. The service is further described in [13].

In the Water4Cities project we collect three different data sets from Ljubljana (Slovenia) and Skiathos (Greece): groundwater information, pump sensor data from Skiathos and weather data. Groundwater dataset contains data from 518 stations comprising 28 regions which measure groundwater levels. Data are collected since 1960 (with median frequency of one day); however, there are some stations that started operating later, or operated intermittently, which means some data are not available.

Weather data are refreshed once per day and contain temperatures (daily average, minima and maxima), location data, precipitation, snow blanket, new snow blanket, cloud cover and sun duration. Data from Ljubljana are available from 2010.

### 2.2. Modeling Approach

We are training a model that will be able to predict continuous groundwater level values based on a set of related attributes (weather data, available historic values of groundwater levels, etc.). We have collected data that combine all the related attributes and target value (groundwater level). Such a problem belongs to the field of supervised learning, more specifically – it is a regression problem.

There is a plethora of algorithms, capable of such a task; however, it is important that our experiments are designed in a way that will give useful and accurate results. Our initial experiments aimed at predicting absolute value of groundwater levels. This proved to be an inefficiently defined problem, since absolute water level depends strongly on long-term historic processes, which we cannot easily grasp with limited attribute vectors. We have, therefore, set our target value not to

absolute groundwater level, but rather to the changes in groundwater levels. For each day we are trying to predict whether and how much the groundwater levels will rise or drop.

### 2.3. Data-Driven Modelling Algorithms

Groundwater level prediction is a regression problem. Based on available data (i.e., weather data, weather predictions, people-behavior model prediction, etc.) we are trying to generate the best possible continuous predictions for groundwater level change on a particular day. Data-driven algorithms use past data to learn the best approximation of an underlying process behind a particular phenomenon.

#### 2.3.1. Linear Regression

One of the oldest and widely used methods is linear regression [14]. We are solving the following equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where  $\mathbf{y}$  is a vector of observed target values (in our case groundwater level or groundwater level change),  $\mathbf{X}$  is a matrix that contains rows of independent variables—attributes (in our case weather, historic or other data),  $\boldsymbol{\beta}$  is a parameter vector we are trying to learn and  $\boldsymbol{\epsilon}$  is a vector of errors. The goal is to learn parameter vector  $\boldsymbol{\beta}$  in order to minimize errors on the validation set. There are many techniques for parameter estimation, the most well-known are based on least squares estimation.

#### 2.3.2. Regression Trees, Random Forests and Gradient Boosting

Regression trees are an algorithm based on decision trees. When learning, we segment the attribute space into many different subspaces, where each particular subspace represented by a tree leaf has a value, which might be obtained simply by averaging all the samples from training set that belong to that leaf or by introducing another model (often linear regression) at each final node.

Regression trees work well in ensembles. Breiman [15] suggested to use ensembles of regression trees to improve prediction accuracy. Each regression tree is trained with a particular sub-sample of a data set (different attributes and different samples from original dataset are used). Final value is given by an average over the whole ensemble. The algorithm is popular as it is fast and easy to parallelize and has proven very successful in numerous applications in environmental data-driven modeling (and elsewhere).

Gradient boosting [16] also utilizes an ensemble of weak learners (usually trees) to provide final prediction, but it stacks them additively. In the first stage the algorithm approximates target values. In each subsequent stage the algorithm approximates pseudo-residuals (loss function differentials) from previous stage. An example of a simple loss function would be

$$L(y, F_m(x)) = \frac{1}{2}(y - F_m(x))^2, \tag{2}$$

where  $y$  is true value and  $F_m(x)$  is the model prediction after  $m$ -th stage. Model prediction  $F_m(x)$  combines all the weak learners' results. Gradient boosting is the state of the art method in various fields (i.e., in particle physics).

### 2.4. Evaluation Methods

We evaluate goodness of fit with root mean squared error (RMSE) and coefficient of determination ( $R^2$ ) metrics. We define them as follows:

$$R^2 = 1 - \frac{\sum_{t=1}^n (f_t - \bar{y})^2}{\sum_{t=1}^n (y_t - \bar{y})^2} \tag{2}$$

$$RMSE = \frac{1}{n} \sum_{t=1}^n (y_t - f_t) \tag{3}$$

$f_t$  represents prediction and  $y_t$  true value at time  $t$ ,  $\bar{y}$  is average true value on the dataset,  $n$  is the number of data samples.

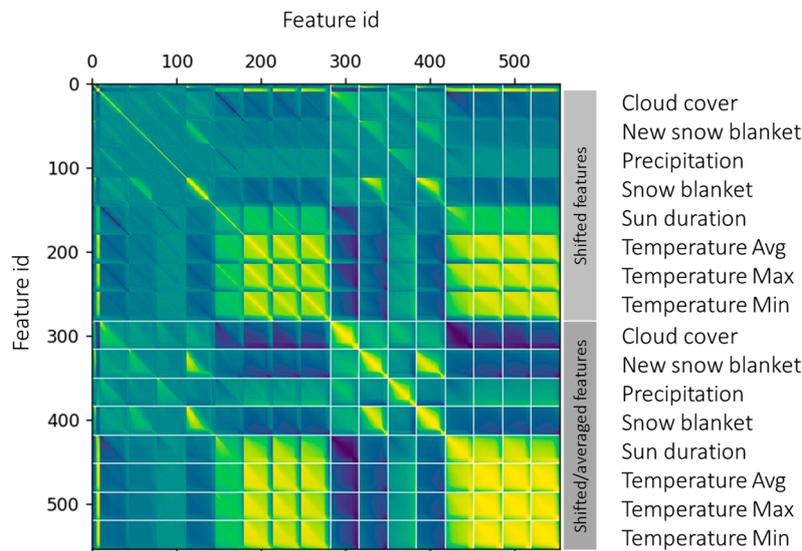
RMSE is a metric that can be easily interpreted in the scope of the use case as it is in the units of the target value. Its main disadvantage is that it cannot be compared across datasets.  $R^2$  is invariant to the amplitude of the target value and can be compared in different scenarios.

We report performance of our algorithms by using 3-fold cross validation.

### 3. Results

#### 3.1. Exploratory Data Analysis

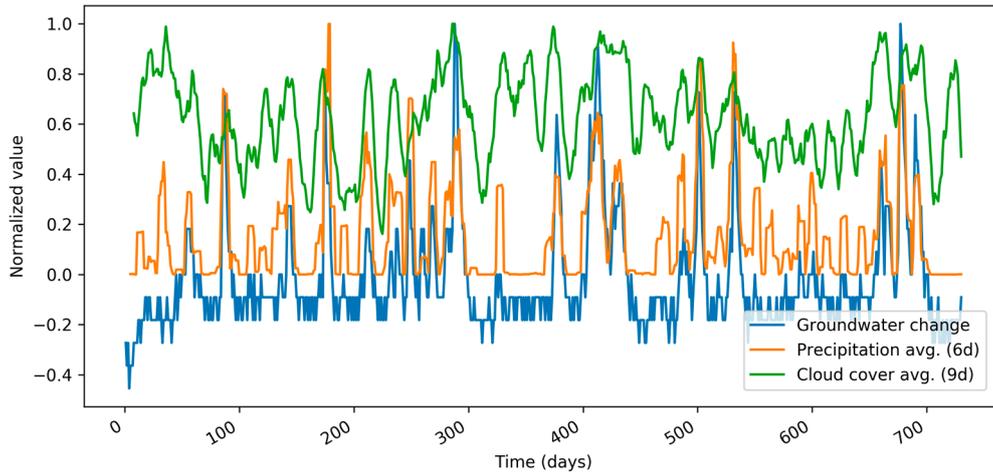
Initial part of the data mining process has been dedicated to feature engineering (engineering of new attributes). We have speculated that different weather phenomena are responsible for groundwater level dynamics: mainly precipitation, snow melting and weather in the whole study area. We have used raw values of all weather attributes (see Figure 1) and different derivatives. Firstly, we introduced time delay in the attribute as we speculated, that it takes time before surface water recharges the aquifer and secondly, we speculated that sums (or averages in the context of data-mining attributes) over multiple past days might play a significant role. Figure 1 depicts the correlation matrix of all 544 generated attributes.



**Figure 1.** Figure represents correlation matrix of 544 original and generated attributes. Lighter color depicts positive correlation, darker color negative correlation. Each weather attribute has been transformed (with various shifts from 1 to 100 days) and then also with averages. Matrix is roughly divided on  $2 \times 2$  fields, each field describing correlations between similar attributes (all 4 regions look similar). In each field we see a significant light block at the bottom right corner, which shows that sun duration and daily temperatures (average and extremes) are highly correlated to one another, even more so, when we take into account averaging over multiple days.

Correlation matrix can be used in two ways. Firstly, we can read the correlations of particular attributes with our target variable (groundwater level daily change) and select the most correlated attributes to use in our models. Secondly, we can use it for filtering of highly correlated attributes. Highly correlated attributes will not bring additional knowledge to the model and might worsen model accuracy. Additionally, number of generated attributes is of the same magnitude as number of samples in our datasets, which might lead to overfitting of the models.

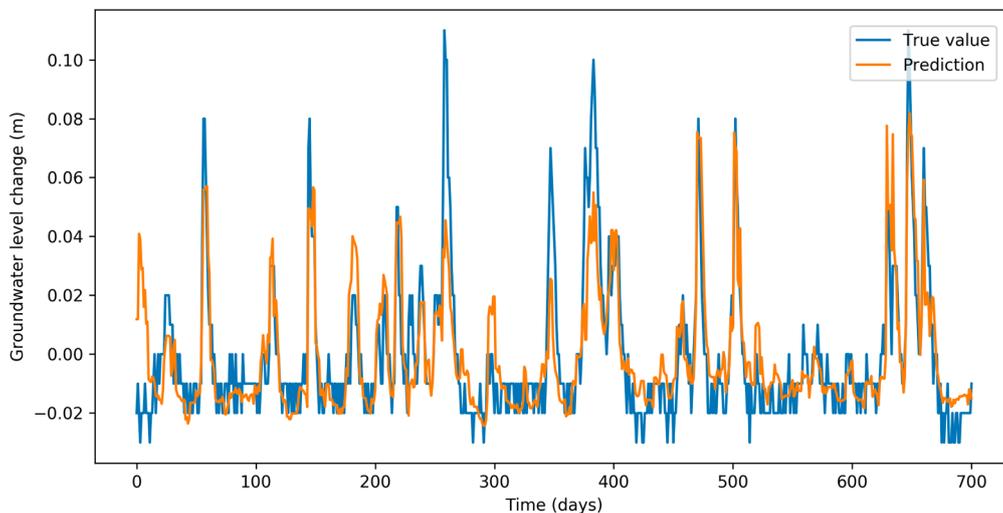
A pair of heterogeneous attributes that are strongly correlated to the target value (groundwater change) are depicted in Figure 2. The attributes are 6-day precipitation average and 9-day cloud cover average.



**Figure 2.** Plot of two distinct highly correlated attributes vs. groundwater change. Especially with precipitation average we can observe that many peaks correspond to peaks in groundwater level change.

### 3.2. Modeling Results

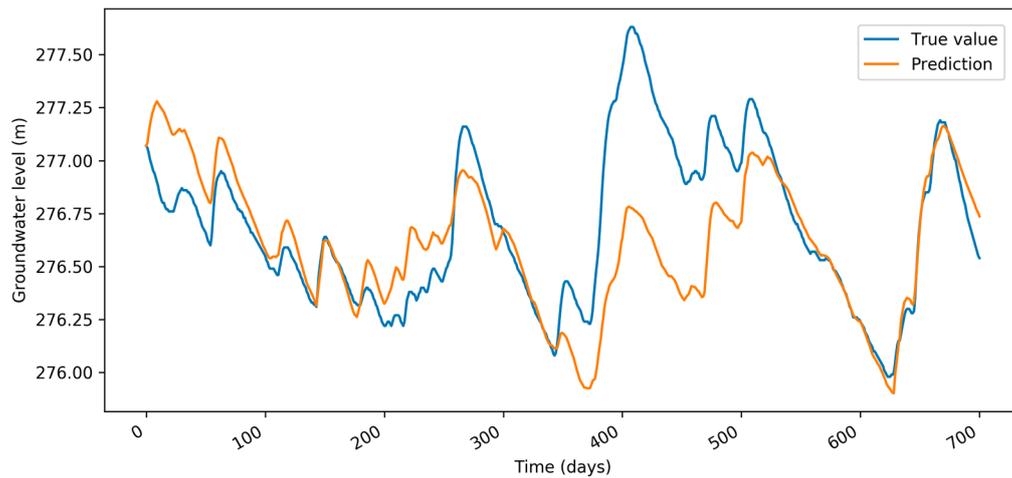
Final feature vectors have been composed of candidates from top-30 correlated attributes. 13 best and mutually least correlated features (according to Figure 1) have been selected. Modeling results are depicted in Figures 3 and 4. Figure 3 depicts comparison between groundwater level change (blue) and its prediction (orange). Visually, we can identify very good fit between the two, which shows that the model explains most of the fluctuations. We can observe that some extreme peaks in groundwater changes are not explained by our features. Further work is needed to identify those.



**Figure 3.** True and predicted groundwater level change based on gradient boosting.

Figure 4 depicts comparison between absolute groundwater level based on our model (orange) and true values (blue). Absolute values are calculated based on summing up groundwater level change predictions over time. We can see that the shape of the orange line reflects dynamics in the blue line, but absolute values might include big differences.

Table 1 presents evaluation metrics on the same data-set using 4 different modeling algorithms: linear regression, regression trees, random forests and gradient boosting. Gradient boosting is the superior method.



**Figure 4.** Comparison between true value and prediction of groundwater level based on gradient boosting. Groundwater level has been calculated with the sum of all the changes from the beginning of prediction.

**Table 1.** Modeling results of groundwater level change prediction, based on weather data. 4 different algorithms were used. Each method has been run on a dataset of 5 different nodes from *Ljubljana polje* aquifer.

Algorithm	R <sup>2</sup>	RMSE
Linear regression	0.624	2.23 × 10 <sup>-4</sup>
Decision trees	0.415	3.46 × 10 <sup>-4</sup>
Random forest	0.609	2.31 × 10 <sup>-4</sup>
Gradient boosting	0.644	2.11 × 10 <sup>-4</sup>

#### 4. Discussion

In this work, we have demonstrated that data-driven modeling techniques within water-management domain can yield useful results with models that are computationally and implementation-wise much less complicated than the current state-of-the-art process-driven models. We have conducted experiments on *Ljubljana polje* aquifer groundwater data.

Based on location of the groundwater level sensors we have achieved satisfactory results, however—we have encountered some anomalies, where our models were unable to model the underlying dynamics of the aquifer. Most of the anomalies are the drought between 2011 and 2012, when groundwater level dropped close to an apparent minimal level. As the models have no information about current groundwater level, they were unable to catch this dynamics. Additionally, some sensors placed near the surface water sources seem to fluctuate close to some predetermined level and are not so much dependent on weather itself. All these anomalies could be handled with additional features, better understanding of the aquifer dynamics and more precisely defined use-case scenarios.

Future work includes many interesting research directions. Firstly, a systematic study on a wide variety of water-management use cases should be performed. Different data-driven approaches should be tested with a wide variety of algorithms (SVM regression, kNN, multi-layer perception ...). So far, researchers have used arbitrary machine learning methods (often they have not even compared them to other available methods). The aim of the study would be to identify the best methods within the domain.

As deep learning is becoming more popular in multiple fields, the methods reported in recent water-management literature, should be further investigated and improved. Different neural network architectures should be tested. Experience from other fields (i.e., particle physics), which deal with similar streams, show that we can often improve state-of-the-art results with deep

learning; however, improvement might not be significant and computationally cheaper methods (like random forests or gradient boosting) can achieve competitive results.

Another approach towards computationally less demanding (and thus so called *green*) methods would be the implementation of stream learning techniques. IoT paradigm is expected to flourish within the water-management practice in the years to follow. Such approaches, which would involve a wide range of prediction models, can be beneficial.

In many domains, such as energy management, data-driven approaches have been pushed aside due to different privacy-related issues. Similar tendency is noticed in water management field. Therefore, it is crucial to stimulate relevant stake-holders to use the methodologies in order to improve decision making processes within their respective organizations. Implementation of such methodologies is an important step towards integration of intelligent decision support systems.

Additionally, a thorough evaluation and comparison of process and data-driven models should be conducted, in order to register limitations, advantages and usability of the latter in real-life scenarios.

## 5. Conclusions

In this study we conduct an analysis for forecasting groundwater levels. Four data-driven methodologies are tested; namely, linear regression, decision trees, random forests, and gradient boosting. Prediction of groundwater level change proved to be a better approach in comparison to the prediction of absolute groundwater levels. All models tested are multivariate and the predictors inserted are: cloud cover, snow blanket, new snow blanket, precipitation, sunlight duration, and average, maximum and minimum temperatures. A variable delay between groundwater change and the predictors was introduced to simulate the actual dynamics of aquifer recharge. Cloud cover and precipitation were chosen to better predict the target attribute. Gradient boosting resulted as the best fitting method with  $R^2 = 0.644$  and  $RMSE = 2.11 \times 10^{-4}$ . According to the results, the method catches adequately the trend but is not catching some groundwater level high peaks.

**Author Contributions:** K.K. conceived and designed the experiments and wrote the paper; K.K., M.Č. and M.B. performed the experiments, K.K., P.P. and C.L. prepared part of the related work and put machine learning experiments into water management context; M.S. prepared the data infrastructure, acquired and cleaned the data; D.M. provided additional analysis of the data.

**Acknowledgments:** The work described in this paper has been conducted within the project Water4Cities. This project has received funding from the European Union's Horizon 2020 Research and Innovation Staff Exchange programme under grant agreement number 734409. This paper and the content included in it do not represent the opinion of the European Union, and the European Union is not responsible for any use that might be made of its content.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Solomatine, D.P.; Ostfeld, A. Data-driven modelling: Some past experiences and new approaches. *J. Hydroinform.* **2008**, *10*, 3–22.
2. Shirmohammadi, B.; Vafakhah, M.; Moosavi, V.; Moghaddamnia, A. Application of Several Data-Driven Techniques for Predicting Groundwater Level. *Water Resour. Manag.* **2013**, *27*, 419–432.
3. Rahmati, O.; Pourghasemi, H.R.; Melesse, A.M. Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. *CATENA* **2016**, *137*, 360–372.
4. Sahoo, S.; Russo, T.A.; Elliott, J.; Foster, I. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. *Water Resour. Res.* **2017**, *53*, 3878–3895.
5. Rejec Brancelj, I.; Smrekar, A.; Kladnik, D. Podtalnica Ljubljanskega polja. *Geografija Slovenije* **2005**, *10*, 1–222.
6. Vizintin, G.; Souvent, P.; Veselič, M.; Curk, B.C. Determination of urban groundwater pollution in alluvial aquifer using linked process models considering urban water cycle. *J. Hydrol.* **2009**, *377*, 261–273.

7. Janža, M.; Meglič, P.; Šram, D. *Numerical Hydrological Model*; Tech. Report Income Water Care EU Life Project; European Commission: Brussels, Belgium, 2011.
8. Auersperger, P.; Čenčur Curk, B.; Jamnik, B.A.; Kus, J.; Prestor, J.; Urbanc, J. Dinamika podzemne vode. Podtalnica Ljubljanskega polja, *Geografija Slovenije* **2005**, *10*, 39–61.
9. Vrzel, J.; Ogrinc, N.; Vižintin, G. Data preparation for groundwater modelling—Ljubljansko polje aquifer system. *RMZ-M&G* **2015**, *62*, 167–173.
10. Kranjc, M. *Data from National Monitoring*; Technical report Income Water Care EU Life Project; European Commission: Brussels, Belgium, 2011.
11. Janža, M. Modelling heterogeneity of Ljubljana polje aquifer using Markov chain and geostatistics. *Geologija* **2009**, *52/2*, 233–240.
12. Pavlič, M.U. *Geološko-Geomehanski Model Zgradbe tal na Območju Mesta Ljubljane*. Ph.D. Dissertation. University of Ljubljana, Ljubljana, Slovenia, 2016.
13. Senožetnik, M.; Herga, Z.; Šubic, T.; Bradeško, L.; Kenda, K.; Klemen, K.; Pergar, P.; Mladenić, D. IoT middleware for water management. In Proceedings of the 3rd EWaS International Conference, Lefkada, Greece, 27–30 June 2018; (accepted).
14. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*, 2nd ed.; Springer: New York, NY, USA, 2017; pp. 43–100.
15. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
16. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2000**, *29*, 1189–1232.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).