Usage of statistical modeling techniques in surface and groundwater level prediction

Klemen Kenda, Jože Peternelj, Nikos Mellios, Dimitris Kofinas, Matej Čerin and Jože Rožanec

ABSTRACT

The paper presents a thorough evaluation of the performance of different statistical modeling techniques in ground- and surface-level prediction scenarios as well as some aspects of the application of data-driven modeling in practice (feature generation, feature selection, heterogeneous data fusion, hyperparameter tuning, and model evaluation). Twenty-one different regression and classification techniques were tested. The results reveal that batch regression techniques are superior to incremental techniques in terms of accuracy and that among them gradient boosting, random forest and linear regression perform best. On the other hand, introduced incremental models are cheaper to build and update and could still yield good enough results for certain large-scale applications.

Key words | data-driven modeling, groundwater level modeling, incremental learning, stream mining, surface water level modeling

Klemen Kenda (corresponding author)
Jože Peternelj
Matej Čerin
Jože Rožanec
Jožef Stefan Institute,
Jamova cesta 39, 1000 Ljubljana,
Slovenia
E-mail: *klemen.kenda@ijs.si*

Klemen Kenda
Matej Čerin
Jože Rožanec
Jožef Stefan International Postgraduate School,
Jamova cesta 39, 1000 Ljubljana,
Slovenia

Nikos Mellios
Dimitris Kofinas
Department of Civil Engineering,
University of Thessaly,
38334 Volos,
Greece

INTRODUCTION

Water data are becoming increasingly accessible and low-cost. Investments in improvement of data acquisition and data transfers have enabled significant growth of knowledge-intensive economies (Washburn *et al.* 2010; Chourabi *et al.* 2012; Di Nardo *et al.* 2015a). However, there is still a great deal of room for improvement, especially compared to the energy or transportation sectors, as indicated by the expected future infrastructure costs by the sector (Laspidou 2014). On the contrary, water as a resource itself is becoming a more expensive commodity. Water utilities worldwide are incorporating – or have already incorporated – the opportunity costs of capital, operation, maintenance, and environmental impacts to the final price under the Polluter-Pays and the User-Pays principles, commonly accepted by the OECD countries (Rogers *et al.* 2002). Digitalization has penetrated most areas of human activity, including major manufacturing facilities, energy markets, health care, and even well-being, while various methodologies on improving resources management and optimizing consumption, usage or exploitation systems have been tested in various settings producing positive results (UNEP 2013). Water management digitalization process is showing great potential for the usage of modern technologies such as the Internet of Things (IoT) and Artificial Intelligence (AI). The latter can operate as a catalyst for investigating, understanding, forecasting, and optimizing water usage, leakage, fraud, and pollutant detection, flooding and damage prevention and protection (Di Nardo *et al.* 2015b).

AI methodologies, especially statistical modeling techniques from the family of machine learning (ML)

doi: 10.2166/aqua.2020.143

algorithms, have proven to successfully complement or even replace the traditional process-based models, usually, requiring much less data preparation and computing time; they are therefore more easily implemented in real-world scenarios (Adamowski *et al.* 2012; Kofinas *et al.* 2014; Tiwari & Adamowski 2014; Mellios *et al.* 2015). ML techniques have also proven to effectively catch patterns that involve complex interdependencies and non-linearities such as those found in the interdisciplinary boundaries of aquatic systems and ecological systems (Recknagel 2001). On the other hand, process-based models are more generalized and provide results that can be applied in broader areas, whereas ML models usually target a specific point in space. One should note that ML models are completely data-driven. This means that expert knowledge is required only to select and transform relevant data and their derivatives into meaningful inputs; and for the conclusive phase of evaluating, validating, and interpreting the outputs. Underlying processes are agnostically modeled by the ML methods. This can be perceived as a benefit or as a drawback, since on the one hand, it constitutes the modeling process practically easier, on the other hand, it deprives some of the interpreting function of building up the causal-effect relations (Sarle 1994; Krause *et al.* 2016). The above reveals a great potential (research and practical) and value for combining ML with process models in a similar way to 'injecting humans-in-the-loop' when modeling with ML, either by interactive training or interactive feature selection (Krause *et al.* 2014, 2016; Amershi *et al.* 2015).

During the last decades, scientific literature is overwhelmed by laboratory tests of ML methodologies, as the aforementioned benefits of such techniques have attracted researchers (Maier & Dandy 2000). Often, data preprocessing (including cleaning and data fusion) in such applications is done manually offline. However, transferring the ML applications to real-world scenarios would require automated data processing pipeline from the data source (sensor or web resource, e.g. for weather and weather forecast data) to the final user.

IoT and other sensor data itself are useful and can be applied to many scenarios, however – much better results can be obtained when using multiple interconnected data sources (Manyika *et al.* 2015). For example, predicting groundwater levels in the future will benefit significantly from weather and weather forecast data and perhaps in some special cases from the water withdrawal and population modeling data. This means that multiple heterogeneous data sources have to be combined in real time to achieve the best possible results.

This work, which extends (Kenda *et al.* 2018a), presents a comprehensive survey of standard statistical modeling techniques (ML models) on the use-cases of groundwater and surface water level prediction. To the best of our knowledge, in this paper, another family of ML techniques is introduced to water domain: incremental learning (sometimes referred to also as stream mining). This is specifically suitable for learning from continually generated sensor data as the models are updated with each subsequent measurement without extensive use of computer resources as in traditional batch methods, which require re-learning on the whole historical dataset. Savings are obtained in data storage, computing power and time.

Slovenia is a country with a dense hydrographical network, with a great difference in the amount of precipitation between areas in the east and the west, with areas of regular or occasional flooding or drought and a positive balance between incoming and outcoming waters. The population density and its related pressure on the aquatic environment also differ. Water landscape is affected by the anticipated climate change (Kanakoudis *et al.* 2016), which is causing longer lasting spring and summer droughts as well as less and more imbalanced precipitations. Despite an overall favorable water balance, shortages can be expected in 15% of country's surface area, mostly in the north-eastern part (https://www.arso.gov.si/en/soer/freshwater.html), as well as floods on another 15% of the territory. Since 1992, seven summer droughts have hit agriculture. In 2003, 2.4% of population required water to be supplied with a tanker.

Accurate groundwater and surface water level short-term predictions allow to better understand its dynamics and convey information about coming extreme events; identify factors that affect water consumption as well as optimize operating schedules of related infrastructure (Adamowski *et al.* 2012). This information also allows to better plan in a context of greater water scarcity (Griffin & Chang 1990) and manages the resource in new ways (e.g. by establishing dynamic pricing (Arbués *et al.* 2003)) in order to increase sustainability.

This analysis aims to introduce a handful tool for groundwater level assessment and forecasting. Such a tool could be used for short-term and mid-term water management. Regarding short-term forecasting, it is crucial to foresee on time a coming extreme event, such as a flood or a shortage crisis. These two extremes may cause a series of malfunctions that create problems in terms of well-being. A shortage crisis, when the aquifer level is too low, affects the urban water supply if the aquifer is used as a reservoir. Regarding mid-term forecasting, such a tool can facilitate the optimal planning of water resources management especially when there are conflicting needs and the potential of multiple sources. End-users of the results or of a product tool of this analysis could be a municipality, a water utility, urban designers, industrial, and agricultural sectors, who could benefit by the easiness of such black-box applications that do not require specific, advanced hydrological knowledge.

The main contributions of this paper are as follows:

1. Comparison of 21 different statistical modeling techniques applied to surface and groundwater levels forecasting at different time horizons.
2. Introduction of incremental learning techniques to modeling of surface and groundwater levels.
3. Usage of efficient and automatic feature selection techniques in surface and groundwater modeling.

METHODS

Statistical modeling techniques

As opposed to traditionally used process-based models, data-driven models rely solely on data. The underlying dynamics of a water system is modeled latently. The model is being learned from the data itself and usually does not require external domain knowledge. Domain knowledge can lead to a significant increase of the model accuracy; however, it is introduced into the model through the appropriate selection of data sources and through appropriate transformation of the data (e.g. relevant non-linear combinations of available sensor data help significantly when trying to improve linear models).

Machine learning is a subfield of the wider AI field. The discipline has been blooming since mid-1970s and has provided widely used solutions such as ML translation, typing assistant, spam mail identification, and image recognition (Hastie *et al.* 2009).

In water management, ML has been used for predicting various key variables in water systems such as groundwater levels (Nie *et al.* 2016; Jeong & Park 2019), urban water demand in multiple scales from household to residential (Al-Qunaibet & Johnston 1985; Griffin & Chang 1990; Oyebode 2019), urban water consumption behavior (Ioannou *et al.* 2017), anomaly detection, such as fraud incidents (García Valverde *et al.* 2015; Candelieri 2017), leakage in water distribution networks (Di Nardo *et al.* 2015a, 2015b), and stratification in reservoirs (Soleimani *et al.* 2019). Often ML research in water management is focused on a particular method, the selection of which is not necessarily justified. There are few research papers, which really investigate a wider variety of algorithms and even among these, very few address the fact, that the usage of a particular modeling algorithm does not influence the final results as much as the appropriate use of contextual data (Hastie *et al.* 2009) and that ensuring proper feature generation and data fusion in real-time (in live, real-world systems) is a great challenge until today (Kenda *et al.* 2019).

The usual ML tasks in environmental data analysis include solving regression and classification problems, which are a part of the family of supervised learning, and clustering, which is part of the family of unsupervised learning algorithms. Supervised learning is performed on labeled data (e.g. groundwater level data, where target values to be modeled are known), whereas unsupervised learning can be used in data, where the target values (e.g. data about users, where the stakeholders would like to discover families of users, which behave approximately the same) are not known. The work reported in this paper tackles regression problems (prediction of numerical values, e.g. of groundwater and surface water levels). These problems were also converted into classification problems (e.g. by dividing groundwater level change into different classes and trying to predict those instead of a continuous value), but those did not yield competitive results.

The most relevant and widely used methods in environmental data-driven modeling nowadays are random forest (Hastie *et al.* 2009), gradient boosting (Friedman 2002), and deep learning (Goodfellow *et al.* 2016), which are

based on simpler building blocks like decision trees and perceptrons (Hastie *et al.* 2009). It should be noted that water management modeling often assumes regression problems, for which deep learning does not exhibit as much power as with other problems (e.g. image recognition, text translation, and similar). According to the nature of a dynamic water system, also linear models like linear regression and support vector machines (SVM classifier or SVC and SVM regressor or SVR) with linear kernel (Hastie *et al.* 2009) can achieve very good results while preserving low-computational cost. Quite often, very simple methods like k-nearest neighbors (Hastie *et al.* 2009) can yield good results.

Incremental learning techniques

Traditional statistical modeling techniques use batches of data points to learn. If the observed system is prone to concept drift (Gama *et al.* 2014), which means that the distribution of the target value is changing through time due to changes in the user behavior or in the environment, frequent re-learning of the models is needed, which is time-consuming (e.g. repeating of model learning step after each day or week). Incremental learning techniques (Bifet 2010) are able to update the existing models. The model, that has been taught on a set of learning examples, can be updated with the next one alone. The update itself is much cheaper than re-learning, and often the incremental learning techniques are aware of the concept drift (e.g. when user behavior is changed to a previously unseen mode due to some external reason and this influences the underlying model), which means that they are able to adapt to the change in the new systems behavior much faster.

Some of the tested methods were the streaming perceptron, Hoeffding trees (Domingos & Hulten 2000), and Hoeffding adaptive trees – HAT (Bifet & Gavaldà 2009). Other methods include recursive linear regression, model trees like FIMT-DD (Ikonomovska *et al.* 2015), incrementally learned neural networks (Zhang *et al.* 2018), and incrementally learned SVMs based on stochastic gradient descent (Bottou 2010). Algorithms like decision trees can be used in ensembles (e.g. bagging), where each tree in the ensemble is fed with a subset of input data (Oza 2005). For classification problems, which are rarer in the water management domain, there are many more methods available; however, there are still not many effective implementations of the state-of-the-art methods.

RESULTS AND DISCUSSION

Experiments were conducted on two different datasets: groundwater and surface water levels in Slovenia (see Table 1). Both cases represent typical regression problems where the modeling task is to predict water level for a certain time period (or prediction horizon) in the future.

Initial experiments have shown that predicting absolute water levels is problematic because the system itself is cumulative. Therefore, reducing the prediction problem to the prediction of daily level differences and not the water levels themselves has proven a good approach. Absolute water levels are finally calculated by the addition/subtraction of predictions from a historical absolute true value. Figure 1 presents the experimental workflow. Green boxes represent data retrieval, blue data manipulation, orange modeling tasks, and yellow model evaluation and results (please refer to the online version of this paper to see this figure in color: http://dx.doi.org/10.2166/aqua.2020.143).

Data

Groundwater and surface water data (see Table 1) have been acquired from an online repository (http://vode.arso.gov.si/hidarhiv/pov_arhiv_tab.php) at the Slovenian environment agency (ARSO). Weather data have been retrieved from DarkSky (https://darksky.net/) web service and ARSO historical weather data repository (http://meteo.arso.gov.si/met/sl/archive/). Sensor data have been thoroughly inspected and only the sensors with data in the period from 2010 to (including) 2017 were selected, for which accurate weather could be retrieved as well. Weather data related to underground water modeling have been retrieved

Table 1 | Experimental datasets include 24 time series

Id	Name	Selected sensors	Availability	Frequency
1	Groundwater levels	2	2010–2017	1/day
2	Surface water levels	22	2010–2017	1/day

Two for groundwater levels in the Ljubljana region and 22 for surface water levels in Slovenia.

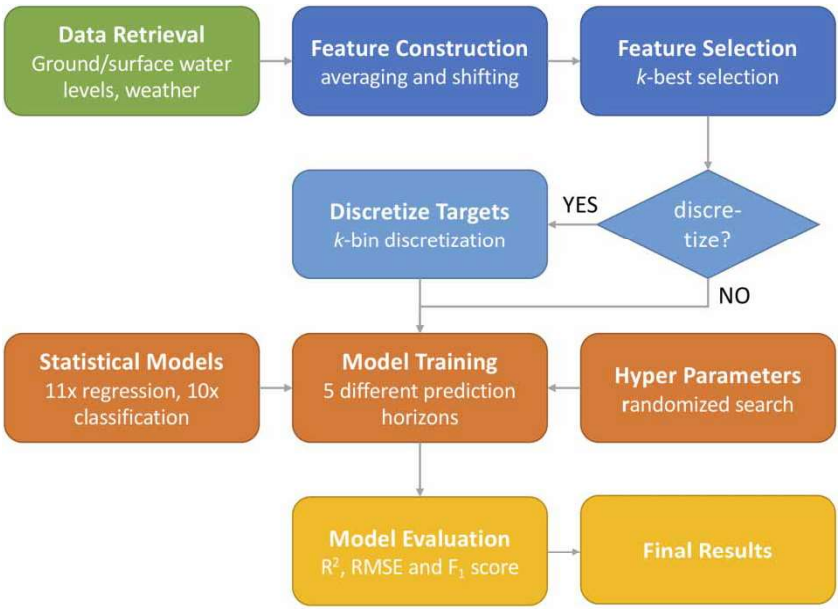

Figure 1 | The workflow of the data-driven approach includes data acquisition, feature generation and selection, modeling and evaluation tasks.

for numerous spatial points in the aquifer, based on the locations of the target sensors, while the weather data for surface water modeling have been retrieved from a single location in the vicinity of the water spring for each selected watercourse.

Although available datasets are larger, only the sensors with clean data that is available throughout the selected period have been chosen for the experiments. Sensors with missing data and with missing or corrupted contextual data (weather) were excluded from the experiments.

The used groundwater levels dataset has been studied extensively (Kenda *et al.* 2018b). The groundwater levels in the aquifer were modeled with linear regression of the values of nearby sensors. The study exposed that the majority of the sensors are highly correlated and could be modeled with extremely high accuracy ($R^2 > 0.995$), while the minority of much less correlated sensors could still be modeled with $R^2 > 0.83$. Due to transitive properties of the operations, the presented methodology could be extended to other sensors and comparable results could be expected.

Feature construction and selection

Statistical models require not only the raw data but also derived features, which reflect a certain physical process that are influencing water level changes. When building additional features, the raw hourly weather forecast data were used, consisting of precipitation probability, precipitation intensity, precipitation type, temperature, cloud cover, dew point, humidity, pressure, and daytime. From these, daily averages, minima and maxima were calculated, producing 24 distinct features which were then analyzed with a correlation matrix (see Figure 2).

A correlation matrix can be used in two ways. Firstly, the correlations of particular attributes with the target variable can be read (water level daily change), and the most correlated attributes can be selected to be used in the models. Secondly, it can be used for filtering of highly correlated attributes. Highly correlated attributes will not bring additional knowledge to the model and might worsen model accuracy.

For example, in this case, pressure turned out to be uncorrelated to the target value and has therefore been removed from the initial set of features. Dew point, on the other hand, has shown a very strong positive correlation to temperature and has therefore also been removed.

The remaining 18 initial features were used to construct additional derivatives by introducing time delays (shifts) and averages over multiple past days. The idea behind this comes from the intuition that groundwater and surface

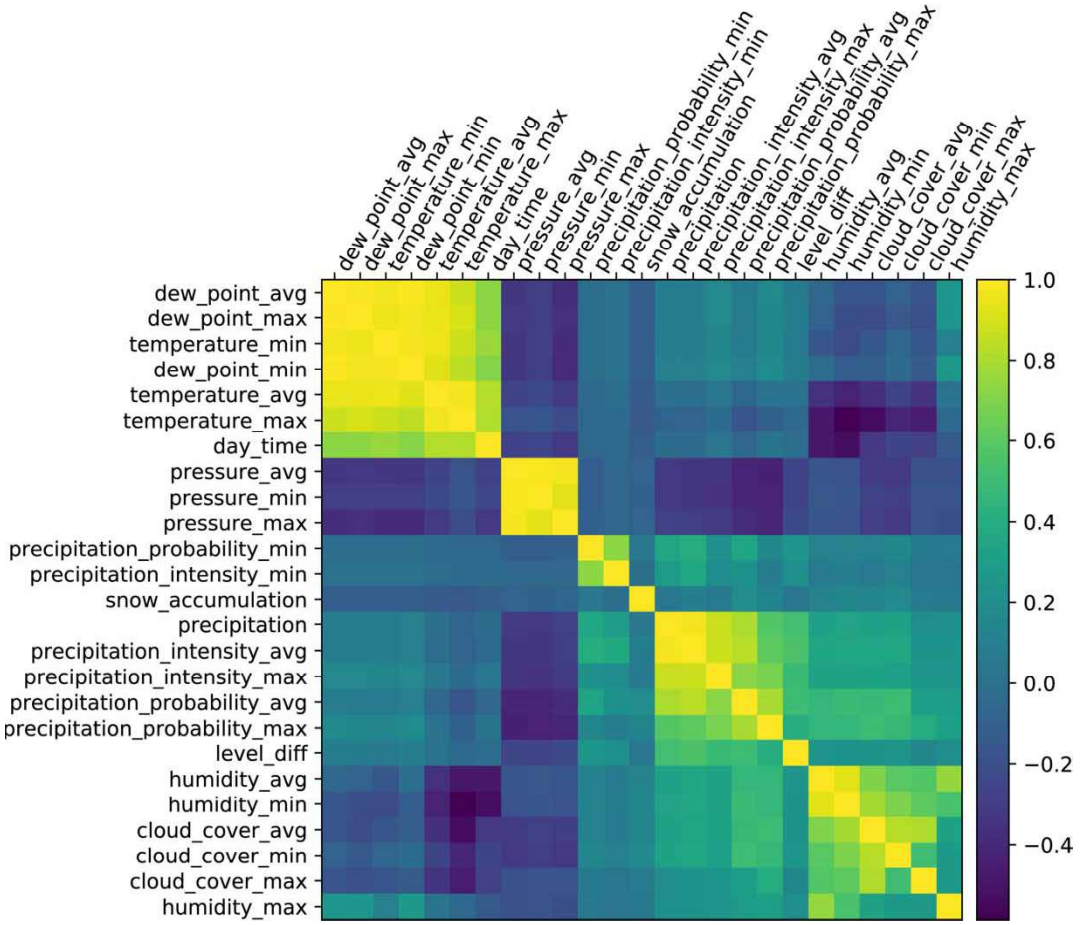

Figure 2 | Correlation matrix of the target value (level_diff) and 24 base features calculated from hourly weather forecasts.

water level responses to the weather changes have some hysteresis.

Since weather forecasts are always limited to finite prediction horizon, the models would always be limited to the same prediction horizon in real-life circumstances. Therefore, it seemed appropriate to introduce the same limitation to the experiments. Specifically, in the majority of the tests, the prediction horizon was set to 3 days. With this limitation in place, it was possible to include another set of features derived from the past water level values, which were constructed with a combination of delays and averages in a similar way as the aforementioned weather feature derivatives.

After the addition of all the feature derivatives, a total of 3,890 features were available. With approximately 3,000 samples of water level measurements in the datasets and the number of features in the same order of magnitude, further steps to reduce the size of the feature vectors were necessary to avoid overfitting. Feature selection (Liu & Motoda 2007) can significantly reduce the number of features without sacrificing the expressivity of the model. In water-related scenarios, one often encounters problems, where the number of measurements of a time series that is modeled is not very high (e.g. one measurement per day in a period of a couple of years). Many methods exist that provide efficient feature selection.

The feature selection approach is based on the selection of a fixed number of top-ranked features, where the *F*-value between features and target values is used as a ranking score. The whole procedure is done in three steps. Firstly, the correlation between each feature and target value is calculated for all training samples as defined in the

following equation:

$$c_i = \frac{\sum_{j=1}^{n} (x_{i,j} - x_{i,\text{mean}})(y_i - y_{\text{mean}})}{\sum_{j=1}^{n} x_{i,\text{std}} y_{\text{std}}}$$

where c_i is correlation of i-th feature to the target value, n is the number of training samples, $x_{i,j}$ is the value of i-th feature in the j-th sample, $x_{i,\text{mean}}$ is the mean value of i-th feature over all samples as defined in the following equation:

$$x_{i,\text{mean}} = \frac{\sum_{j=1}^{n} x_{i,j}}{n}$$

y_{mean} is the mean value of the target value over all samples as defined in the following equation:

$$y_{\text{mean}} = \frac{\sum_{j=1}^{n} y_j}{n}$$

$x_{i,\text{std}}$ is the standard deviation of i-th feature over all samples as defined in the following equation:

$$x_{i,\text{std}} = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (x_{i,j} - x_{i,\text{mean}})^2}$$

and y_{std} is the standard deviation of the target value over all samples as defined in the following equation:

$$y_{\text{std}} = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (y_j - y_{\text{mean}})^2}$$

Secondly, the correlation is converted to F-score as defined in the following equation:

$$f_i = (n-2)\frac{c_i^2}{1 - c_i^2}$$

where f_i is the F-value of i-th feature. Finally, features are sorted by their corresponding F-values, and only first k with the highest score are included in the final selection. In this case, k was experimentally determined to be $k = 30$, which yielded the best results on average. A selection of features is presented in the online Supplementary Appendix.

This whole process of automatic feature extraction and selection was repeated for each dataset used in the experiments.

Evaluation

When evaluating statistical models, the dataset is usually split into three different parts. The first part (training set) is used to train the models, the second part (validation set) to fine-tune the hyper parameters, and the last part (test set) for testing. This is known as the 'hold-out' approach. This approach is dependent on a single split. This limit can be overcome with cross-validation. In this method, the dataset is divided into N parts, each part is used once for testing, once for validation, and $N - 2$ times for training. With such an approach, the random effects of arbitrary train–validation–test splitting are reduced. However, cross-validation assumes that the samples are independent, which is not the case for time series such as water levels due to the causality and autocorrelation of nearby samples. Therefore, an evaluation of the model is only possible on the basis of 'future' observations. A k-fold time series split methodology was used, which is a variant of the k-fold cross-validation, but without including future data in the training set.

With this evaluation methodology, several evaluation criteria are eligible for use when evaluating regression models. The most common are the coefficient of determination (R^2), root-mean-squared error (RMSE), and mean absolute percentage error (MAPE). In the experiments, no major differences between the different scores for the models and data were observed; however, in order to ensure comparability with other experiments and datasets, a measure, that is invariant to data offset and amplitude, was chosen. R^2 preserves both (and has therefore been chosen), while RMSE is sensitive to amplitude, and MAPE is to the offset of data. R^2 has been chosen as the most suitable evaluation metrics for the experiments.

R^2 is defined as $R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$ where y_i is the i-th target value, $\bar{y}$ is the average target value, and f_i is the predicted value.

Note: R^2 is calculated for level differences, which is a rapidly changing time series. This means that the R^2 for

the actual water level is much higher. For example, the highest coefficient of determination for surface water levels was 0.572 for level differences, while for actual surface water levels, it was above 0.93.

Experiments

The evaluation of the experimental results is divided into two sections. The first section is dedicated to the modeling of groundwater levels and the second to the modeling of surface water levels. Each section presents the following results: illustrative prediction results for water level differences and for water levels with multiple prediction horizons, comparison of the accuracy of different methods, and qualitative results on the importance of model features that can be further analyzed by domain experts.

Groundwater level modeling

The experiments were conducted with a set of 11 regression and 10 classification modeling methods. When using multiclass classification methods, the discretization of the target space (daily water level differences) into eight separate classes was used. The feature space normalization was used where necessary (for support vector machines, multilayer perceptron neural network – MLP, k-nearest neighbors, and perceptron). The implementation of the statistical models from scikit-learn (batch learning) and scikit-multiflow (incremental learning) libraries for Python was used.

Illustrative results for the prediction of level differences are shown in Figure 3, and cumulative results for water levels are shown in Figure 4. The overall experimental results are listed in Table 2 and depicted in Figure 5.

Figure 3 depicts the actual results of the prediction experiments. Level differences are calculated for 3 days in advance. The statistical model can predict major changes in groundwater levels with fairly good accuracy (in terms of start, duration, and amplitude). The modeling results are converted into water level predictions as shown in Figure 4.

Figure 4 illustrates the characteristics of different types of statistical models for groundwater level prediction. The best regression models, the best streaming regression models, and the most illustrative classification-based

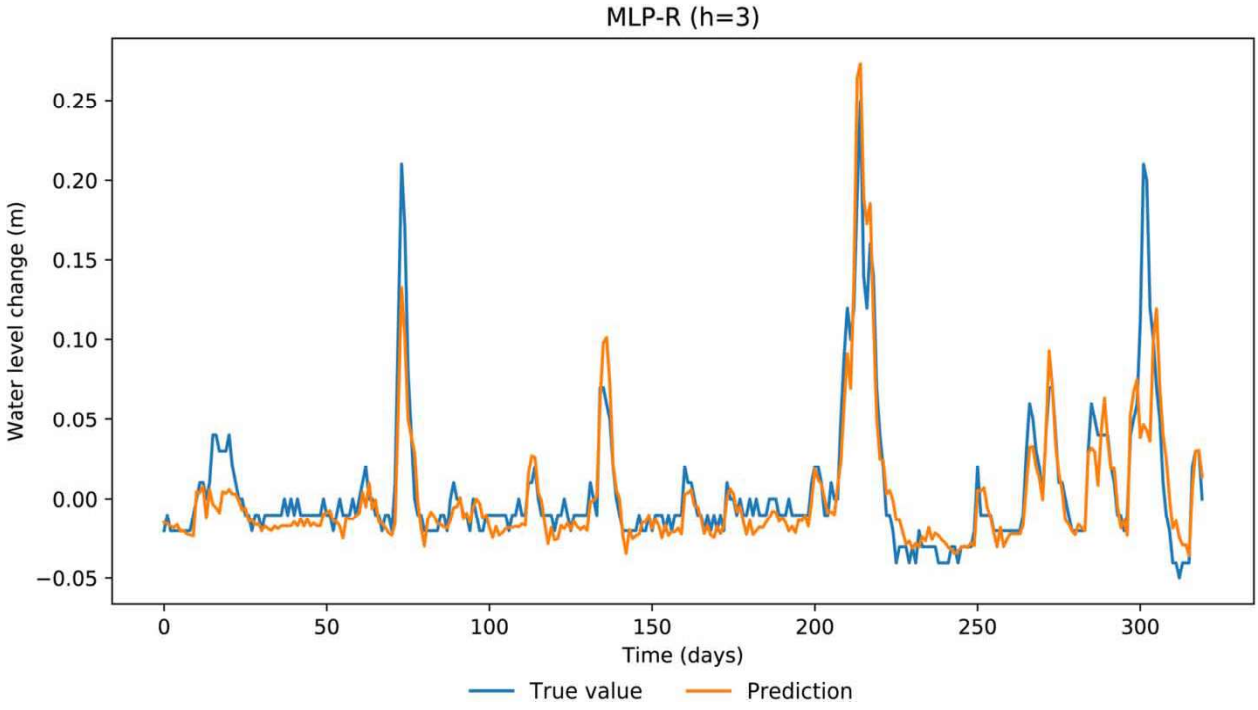

Figure 3 | Prediction of groundwater level difference (change) for MLP (multilayer perceptron neural network) regressor for 3-day prediction horizon.

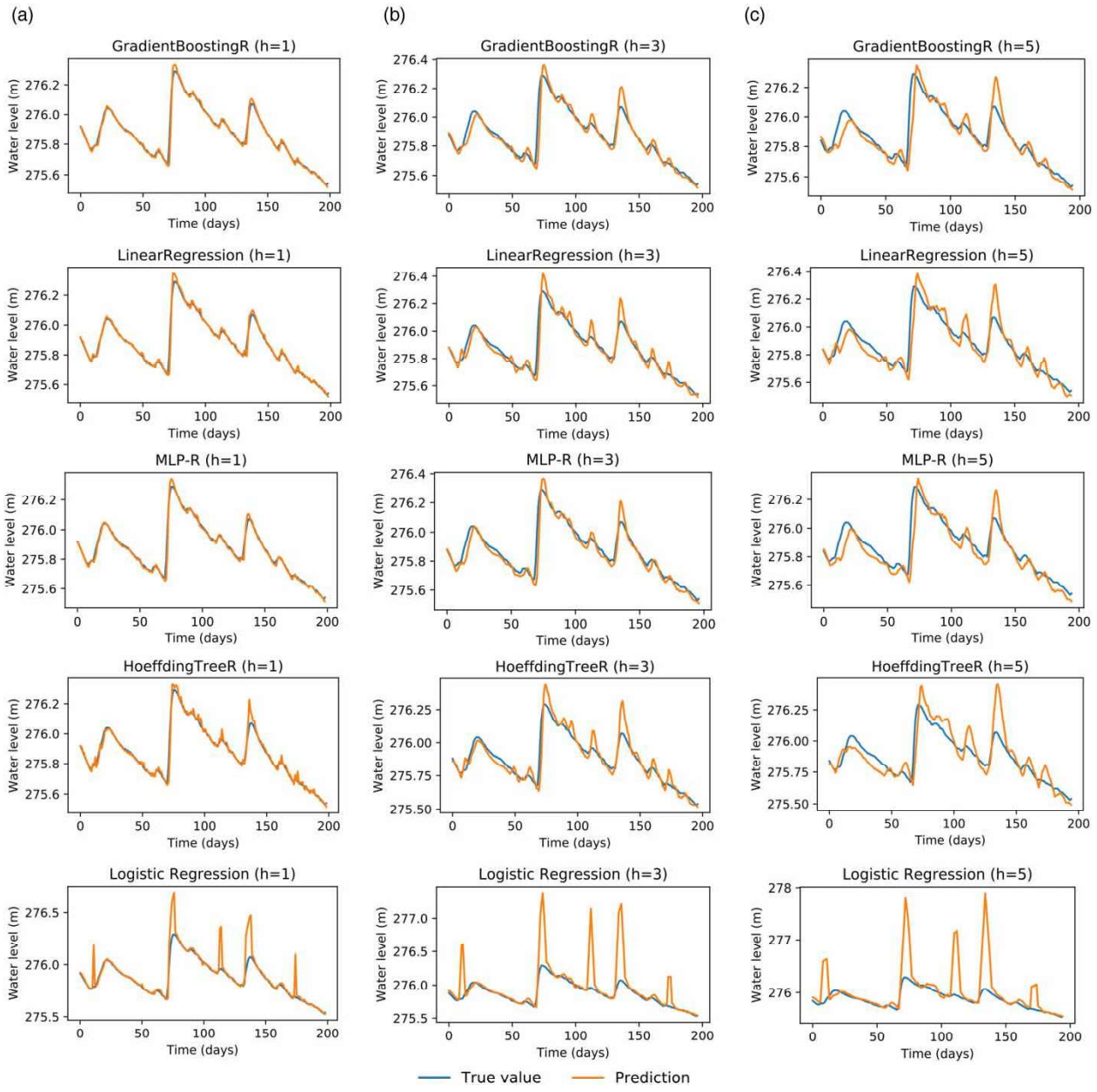

Figure 4 | Illustrative results of different prediction models for groundwater levels. Different algorithms are depicted in rows, different prediction horizons: (a) 1 day ahead, (b) 3 days ahead, and (c) 5 days ahead are shown in columns.

models are included in the figure. The majority of models overshoots larger level changes. This means that in historical data in some cases includes bigger level changes under similar conditions. This could, for example, indicate higher water withdrawal or a change in the dynamics of the groundwater system. Prediction accuracy decreases with the prediction horizon and that the models sometimes miss the beginning of a change (with longer prediction horizons). The only model that correctly records the beginning of the water level change properly even for a 5-day prediction horizon is the classification model. This means that although the accuracy of the model is poor, the classification

Table 2 | Averaged modeling results for groundwater levels for five different prediction horizons (from 1 to 5 days ahead)

	1 day ahead		2 days ahead		3 days ahead		4 days ahead		5 days ahead			
Method	R^2	σ	R^2	σ	R^2	σ	R^2	σ	R^2	σ	t_t (ms)	t_p (ms)
LinearRegression	0.834	0.037	0.708	0.063	0.661	0.066	0.638	0.063	**0.641**	0.07	1.9	0.2
DecisionTreeR	0.677	0.146	0.61	0.111	0.545	0.093	0.541	0.061	0.543	0.059	3.3	0.1
RandomForestR	**0.842**	0.031	**0.726**	0.108	**0.669**	0.08	0.644	0.067	0.600	0.054	1,047	8.5
GradientBoostingR	**0.849**	0.037	**0.775**	0.052	**0.732**	0.071	**0.690**	0.081	**0.655**	0.09	322.6	0.5
PLSRegression	0.726	0.052	0.65	0.073	0.639	0.076	0.639	0.076	0.639	0.077	2.9	<0.1
ExtraTreeR	0.677	0.146	0.61	0.111	0.545	0.093	0.541	0.061	0.543	0.059	3.3	<0.1
SVR	− 0.137	0.41	− 0.16	0.356	− 0.113	0.325	− 0.069	0.317	− 0.086	0.324	2.5	0.4
MLP-R	0.825	0.045	0.691	0.084	0.659	0.09	**0.646**	0.092	**0.641**	0.082	210.2	1.3
KNeighborsR	0.747	0.053	0.661	0.089	0.631	0.102	0.618	0.112	0.610	0.114	5.9	8.0
HoeffdingTreeR	0.495	0.164	0.513	0.127	0.518	0.144	0.493	0.139	0.482	0.13	3,347.1	28.1
HAT-R	0.506	0.176	0.513	0.127	0.518	0.144	0.493	0.139	0.482	0.13	3,722.2	28.5
LogisticRegression	0.485	0.153	0.408	0.131	0.395	0.146	0.376	0.179	0.370	0.16	83.3	0.2
DecisionTreeC	0.506	0.141	0.395	0.116	0.342	0.183	0.365	0.203	0.377	0.183	4.9	<0.1
ExtraTreeC	0.306	0.09	0.336	0.113	0.24	0.162	0.332	0.058	0.256	0.258	2.4	0.1
RandomForestC	0.554	0.108	0.489	0.178	0.481	0.19	0.498	0.176	0.470	0.186	279.8	9.8
SVC	0.522	0.088	0.413	0.131	0.375	0.158	0.391	0.193	0.400	0.191	135.4	16.9
KNeighborsC	0.530	0.071	0.378	0.143	0.387	0.201	0.357	0.157	0.353	0.16	7.8	15.8
Perceptron	0.435	0.206	0.102	0.388	− 0.007	0.788	0.325	0.291	0.266	0.268	10.5	0.2
GaussianNB	0.472	0.112	0.378	0.123	0.374	0.138	0.362	0.144	0.358	0.148	1.2	0.3
HoeffdingTreeC	0.472	0.112	0.378	0.123	0.374	0.138	0.362	0.144	0.358	0.148	1,054.2	119.5
HAT-C	0.453	0.108	0.371	0.14	0.364	0.153	0.346	0.147	0.353	0.148	1,912.2	120.4

Best results by prediction horizon are bolded. Best classification-based results are underlined.

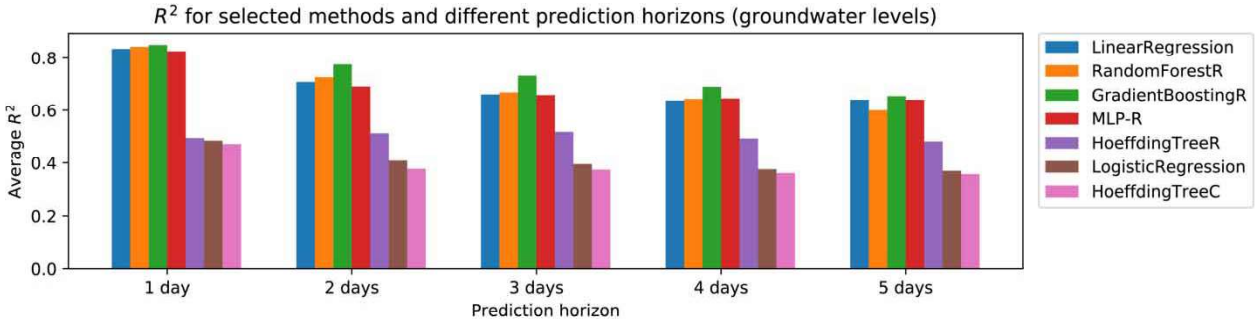

Figure 5 | R^2 of selected statistical models for different prediction horizons (groundwater levels).

could be a good choice for estimating the start of level changes. In addition, the reason for poor classification results can be seen from the figure. During the discretization, the rarely occurring high values of the level changes were grouped into a single bin (which produces a fixed amplitude), although these values are distributed over a bigger interval. For extreme values, finer discretization would lead to better results.

The modeling methods perform as expected. Traditional workhorses perform best (with R^2 scores between 0.6 and

0.85), and incremental learning methods and classifiers yield substantially worse results. Gradient boosting is significantly better than competing methods. As shown in Kenda *et al.* (2018a), linear regression is also a good choice in this setup. However, it is important to note that extensive feature engineering has helped it catch the latent dynamics of the aquifer, while gradient boosting could perform reasonably well even without so many features. The discretization of daily level differences in classes and the use of multi-class classification did not perform well. During discretization some information is lost, which is reflected in the results.

Hoeffding Tree regressor performed best among incremental learning methods. It maintains its performance even with longer prediction horizons and is more competitive in scenarios with a prediction horizon of more than 2 days. The use of incremental learners is most effective in scenarios where the distribution of target values changes over time, which is not the case for groundwater and surface water levels in Slovenia between 2010 and 2017. The methods could be much more effective in modeling the behavior of water consumers.

The main advantage of the usage of incremental learning methods is that they do not have to re-learn from all the data in the training phase. They can only be updated with the latest value and are thus computationally much more efficient. This could be taken advantage of in scenarios where many sensors with fast updates are used (e.g. modeling consumers' behavior in a large city). Batch models should be retrained regularly. The training (t_t) and prediction (t_p) times listed in Table 2 show that linear regression could be the most effective method in practice, since its training and prediction times are among the smallest, while the competitive methods like random forest, gradient boosting, and MLP require significantly more time to (re-)train.

R^2 scores are used as a comparative measure to select the best possible methodology. Comparison of the R^2 scores with other studies is possible on the illustrative level only, since different datasets are based on the aquifers with different internal processes. Currently, no standardized dataset to test different approaches exists for groundwater (and surface water) levels. Illustrative comparison to the recent state of the art (Chen *et al.* 2020) shows that the presented methodology could achieve superior results (R^2 scores for 1 day ahead are by 0.1 larger than the compared results in a 'now-casting' scenario). This could be attributed to extensive feature engineering, higher number of tested

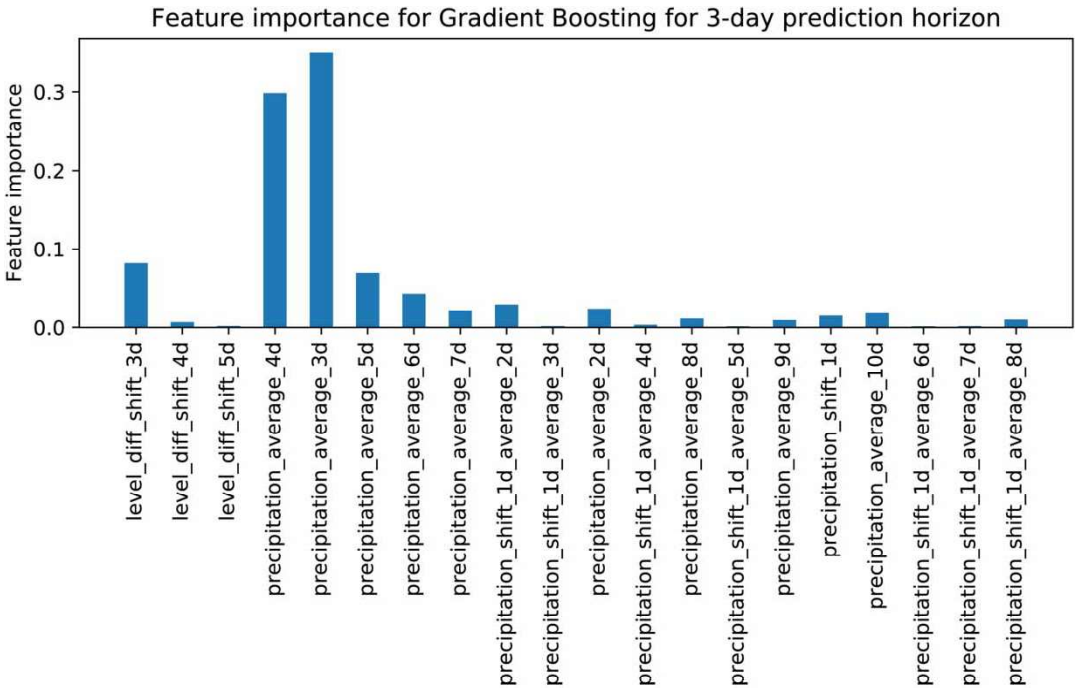

Figure 6 | Feature importance scores (relative) as given by the gradient boosting regression model for 3-day prediction horizon for groundwater level modeling.

methods and intelligent feature selection. The study (Chen *et al.* 2020) also shows that data-driven models give superior results to the traditional process-based models by a high margin (R^2 scores of data-driven models are higher by approximately 0.2).

Most ML algorithms provide a kind of score on the importance of a feature in the statistical models. In linear regression, these are coefficients, and in ensemble tree-based methods, these are importance weights. An example importance weights for a 3-day prediction horizon for gradient boosting is shown in Figure 6. The groundwater level change is influenced by the current trend (last level difference) and different values related to precipitation and its history. Among other weather phenomena, none was selected using the automatic method. The current average precipitation values are the most important ones. Some shifted features also play an important role.

This analysis shows that the automatic feature selection algorithm overlooked some of the seasonally important features, such as data related to snow/snow melting, cloud cover, and similar, which is a consequence of the correlation-based approach. These features could significantly improve the accuracy of the algorithms in the respective season.

Finally, it is worth mentioning that an automatic methodology to produce the reported statistical models was developed. The performance of the automatic methodology could be improved by using a better feature selection algorithm (some have been tested) that would select most informative features based on their modeling performance. A genetic search algorithm across the modeling feature space could be the most efficient. Of course, modeling the water level in a particular well can benefit significantly from the input of the domain expert (what features to use and what additional data is required).

Surface water level modeling

The same methodology was used for modeling the surface water levels as for groundwater. Illustrative results of surface water level differences and levels are depicted in Figures 7 and 8. The results are listed in Table 3. The selected sensor accuracy in terms of R^2 is depicted in Figure 9.

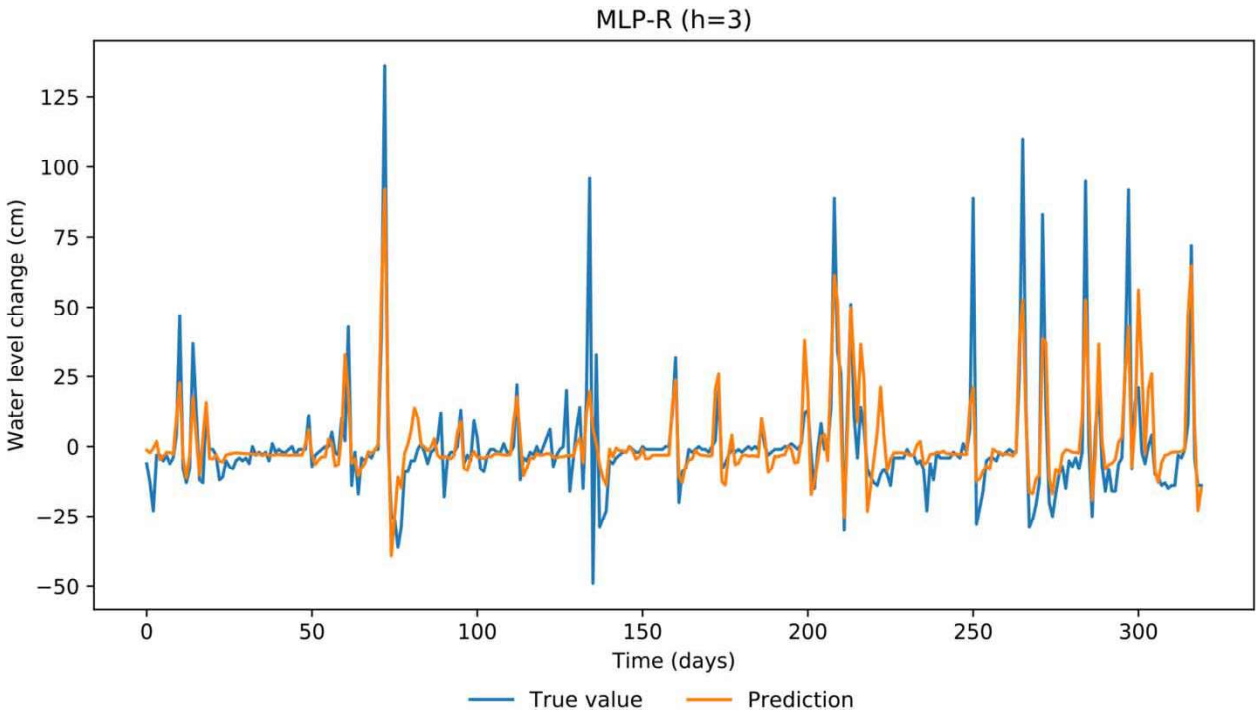

Figure 7 | Prediction of surface water level difference (change) for MLP (multilayer perceptron neural network) regressor for 3-day prediction horizon.

Figure 10 also shows the standard deviation of R^2 scores among 22 sensors for three selected methods.

Comparing Figure 7 to Figure 3 reveals larger discrepancies between the prediction (orange) and the true values (blue) (please refer to the online version of this paper to see these figures in color: http://dx.doi.org/10.2166/aqua.2020.143). For example, it can be observed that the model was unable to follow the dynamics in the real world between days 120 and 150. The fluctuations, even the large ones, are sometimes not reflected in the modeling results. This means that certain mechanisms in nature could not be modeled with the input data. The model is also conservative in estimating the high peaks in the level change. This usually means that similar situations (similar feature

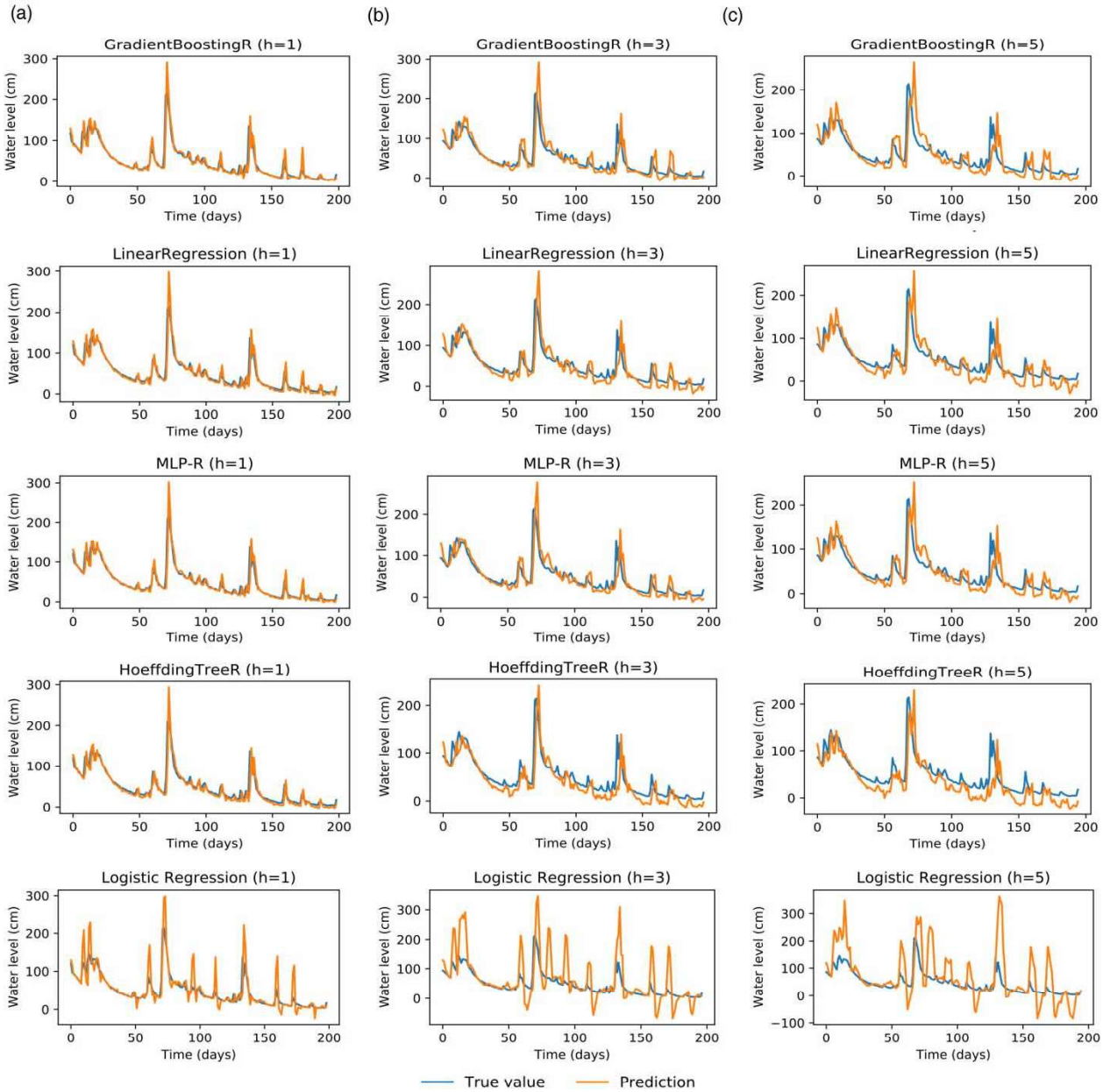

Figure 8 | Illustrative results of different prediction models for surface water levels. Different algorithms are depicted in rows, different prediction horizons: (a) 1 day ahead, (b) 3 days ahead, and (c) 5 days ahead are shown in columns.

Table 3 | Averaged modeling results for surface water levels for five different prediction horizons (from 1 to 5 days ahead)

	1 day ahead		2 days ahead		3 days ahead		4 days ahead		5 days ahead			
Method	R^2	σ	R^2	σ	R^2	σ	R^2	σ	R^2	σ	t_t (ms)	t_p (ms)
LinearRegression	0.553	0.069	0.499	0.071	0.487	0.081	0.484	0.081	0.489	0.078	3.7	0.3
DecisionTreeR	0.350	0.172	0.325	0.161	0.287	0.202	0.294	0.195	0.267	0.302	3.8	0.1
RandomForestR	**0.571**	0.106	**0.514**	0.109	**0.492**	0.109	**0.501**	0.111	**0.502**	0.1	1,194	6.4
GradientBoostingR	**0.572**	0.106	0.500	0.088	0.480	0.108	0.480	0.113	0.482	0.109	399.0	0.4
PLSRegression	0.502	0.065	0.457	0.068	0.450	0.072	0.450	0.071	0.449	0.07	3.0	0.4
ExtraTreeR	0.350	0.172	0.325	0.161	0.287	0.202	0.294	0.195	0.267	0.302	3.7	0.1
SVR	0.154	0.059	0.155	0.059	0.156	0.058	0.156	0.058	0.156	0.058	158.6	18.2
MLP-R	0.538	0.078	**0.512**	0.077	**0.511**	0.083	**0.511**	0.082	**0.509**	0.082	3,014	1.5
KNeighborsR	0.507	0.092	0.458	0.095	0.439	0.111	0.435	0.109	0.436	0.103	8.6	12.7
HoeffdingTreeR	0.312	0.212	0.246	0.203	0.279	0.178	0.284	0.181	0.282	0.18	3,748	39.8
HAT-R	0.312	0.213	0.218	0.28	0.276	0.183	0.281	0.186	0.279	0.184	4,140	40.1
LogisticRegression	0.205	0.162	0.212	0.158	0.207	0.167	0.200	0.159	0.206	0.169	110.6	0.2
DecisionTreeC	− 0.066	0.301	− 0.108	0.353	− 0.096	0.35	− 0.09	0.383	− 0.104	0.35	6.3	0.1
ExtraTreeC	− 0.189	0.447	− 0.181	0.403	− 0.163	0.371	− 0.252	0.479	− 0.192	0.455	2.1	0.1
RandomForestC	0.081	0.241	0.065	0.254	0.069	0.221	0.08	0.224	0.058	0.248	264.5	8.5
SVC	0.083	0.174	0.130	0.178	0.126	0.173	0.124	0.178	0.135	0.172	167.9	20.1
KNeighborsC	0.06	0.18	0.047	0.206	0.029	0.206	0.039	0.207	0.037	0.209	8.8	17.6
Perceptron	− 0.121	0.739	− 0.224	0.732	− 0.279	0.964	− 0.14	0.736	− 0.088	0.622	14.0	0.2
GaussianNB	0.127	0.225	0.143	0.217	0.143	0.219	0.142	0.22	0.139	0.218	1.3	0.4
HoeffdingTreeC	0.097	0.221	0.111	0.216	0.111	0.218	0.109	0.218	0.111	0.216	1,592	154.7
HAT-C	0.099	0.218	0.115	0.209	0.108	0.207	0.108	0.206	0.108	0.208	2,694	161.8

Best results by prediction horizon are bolded. Best classification-based results are underlined.

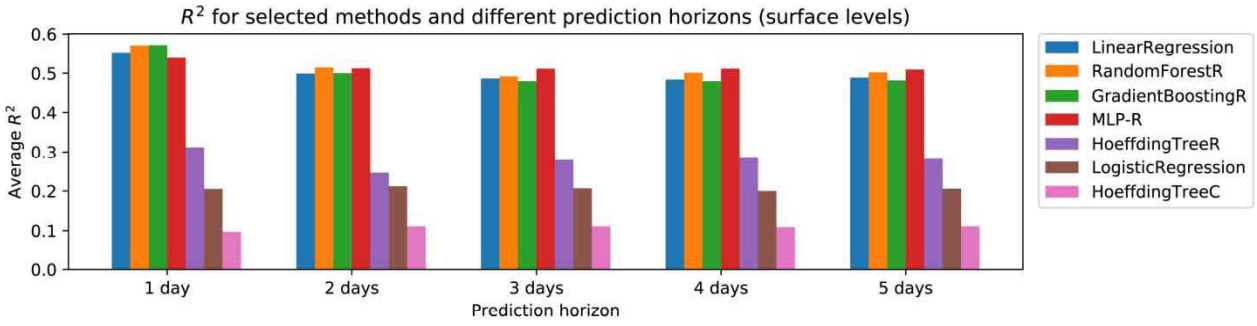

Figure 9 | R^2 of selected statistical models for different prediction horizons (surface water levels).

vectors) also occur in the learning data in cases where the change is smaller and the model learns to predict the value in between.

Compared to Figure 4, Figure 8 shows predictions that are less accurate. Furthermore, it can be concluded that the surface water levels are much less stable than groundwater levels. The changes in water levels are more rapid (the system is more responsive). The water levels are rising but also decreasing much faster. It is much more difficult to accurately predict such behavior.

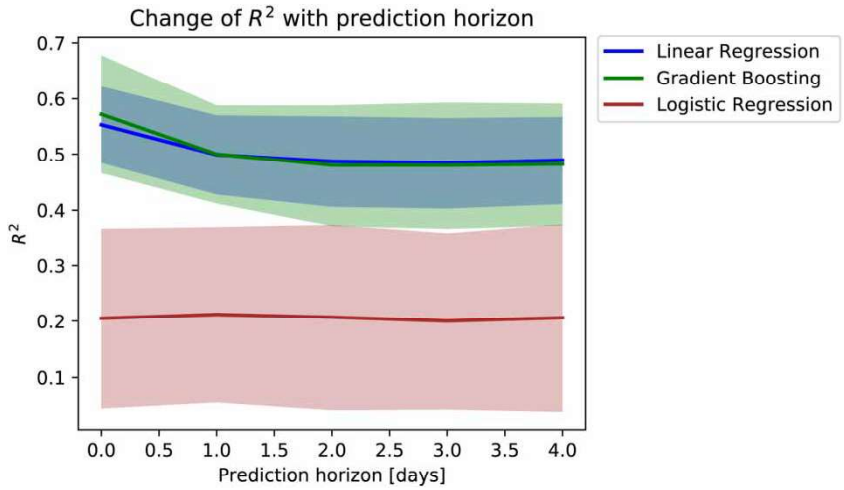

Figure 10 | Change of R^2 and its standard deviation with prediction horizon for linear regression, gradient boosting, and Hoeffding tree regression.

The performance of statistical modeling techniques in this case is similar to the modeling of groundwater levels. R^2 is generally lower than in the case of groundwater, which means that the input data did not include all the features that reflect the system dynamics and that the dynamics of surface water is less stable compared to groundwater. There is also a greater relative discrepancy between the batch regression methods and the streaming and classification methods, which are unsuitable in this scenario.

A further visualization of R^2 is shown in Figure 10 where beside the value of R^2 also the standard deviation of R^2 in 22 experiments (for 22 different stations) is shown. A decrease in accuracy in linear regression and gradient boosting for longer prediction horizons can be observed. It is also evident that linear regression is more stable than gradient boosting because the distribution of results is closer to the mean. The classification-based method (logistic regression) behaves significantly worse and gives unstable results.

Feature importances of gradient boosting regression with a 3-day prediction horizon for a selected surface water level sensor are shown in Figure 11. The precipitation intensity averaged over 2 days is the predominant characteristic that is easy to interpret. It is more important, whether precipitation can infiltrate the ground than how much precipitation there is. Excess water is transported on the surface and raises the surface water levels. Other features are far less important. The precipitation probability and its derivatives represent the other family of important features that influence the model.

CONCLUSIONS

This paper provides a comprehensive overview of the performance of statistical modeling techniques in applications of groundwater and surface water level forecast. Standard batch and incremental ML techniques for regression and classification are included. The latter are used on binned target values of water levels.

Comparison of regression techniques with classification methods on discretized bins reveals that the classification techniques are significantly inferior. An interpretation is that even though classification offers a wide range of ML techniques, the nature of the data is such that no binning is possible without worsening the results. The final performance depends heavily on how well the targets are binned, which is a complex task that must consider the density and distribution of values and is a matter of subjective interpretation. On the other hand, regression techniques can naturally handle the prediction of a target continuum of values, which leads to better results. Nevertheless, the classification techniques are much more successful in determining the starting time of a groundwater level change in longer prediction horizons. In combination with regression

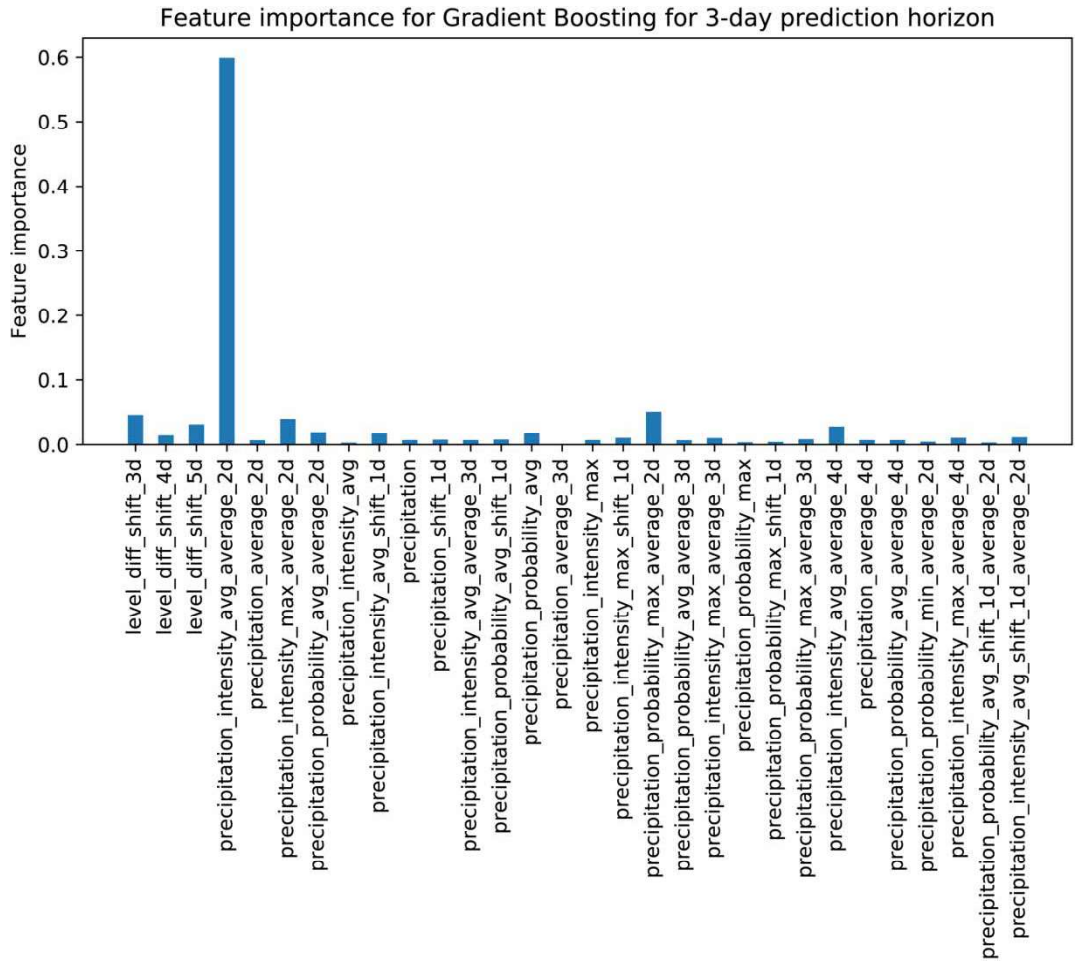

Figure 11 | Feature importance scores (relative) as given by the gradient boosting regression model for 3-day prediction horizon for surface water level modeling.

techniques, this could represent a potentially superior modeling approach.

Analysis of regression methods shows that despite streaming techniques adapt to incoming data and detect concept drift, they are consistently outperformed by their batch counterparts by a significant margin. Among streaming methods, the best results were obtained with the Hoeffding tree regression, which could provide competitive results (in terms of computational performance) in some scenarios. Gradient boosting, multilayer perceptron neural network, random forest regression, and linear regression achieved the best results in both use-cases. The good performance of linear regression can be attributed to the extensive feature engineering and to the nature of the underlying physical models.

Automatic feature engineering and feature selection algorithms, that enrich the water level values with contextual information and later prune it in order to avoid overfitting of ML models, are important contributions of the presented approach.

Finally, this work could be extended in various directions. The main challenge would be to provide a more comprehensive approach to the feature selection. The current approach has explored similarity-based and information gain-based methods; however, wrapper methods are expected to give even better results. Since the datasets are relatively small, the latter approach could find the nearly optimal set of features per sensor/prediction horizon in a reasonable time.

End-users can benefit from the effectiveness, accessibility, simplicity, and speed of the presented modeling solution. In order to put the AI methods into practice, a suitable Big Data architecture should be developed that

can handle automatic data acquisition, transformation, and fusion, as well as the generation of predictions in near real-time. Without these, the modeling results remain in the laboratory.

ACKNOWLEDGEMENTS

This research was funded by the European Union's Horizon 2020 programme project Water4Cities (Research and Innovation Staff Exchange) grant number 734409 and the European Union's Horizon 2020 programme project NAIADES (Innovation Action) grant number 820985.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this paper is available online at https://dx.doi.org/10.2166/aqua.2020.143.

REFERENCES

Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B. & Sliusarieva, A. 2012 Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resources Research* **48** (1), W01528.

Al-Qunaibet, M. H. & Johnston, R. S. 1985 Municipal demand for water in Kuwait: methodological issues and empirical results. *Water Resources Research* **21** (4), 433–438.

Amershi, S., Chickering, M., Drucker, S. M., Lee, B., Simard, P. & Suh, J. 2015 Modeltracker: redesigning performance analysis tools for machine learning. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, pp. 337–346.

Arbués, F., Garcıa-Valiñas, M. Á. & Martınez-Espiñeira, R. 2003 Estimation of residential water demand: a state-of-the-art review. *The Journal of Socio-Economics* **32** (1), 81–102.

Bifet, A. 2010 Adaptive stream mining: pattern learning and mining from evolving data streams. In: *Proceedings of the 2010 Conference on Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams*. IOS Press, Amsterdam, The Netherlands, pp. 1–212.

Bifet, A. & Gavaldà, R. 2009 Adaptive learning from evolving data streams. In: *International Symposium on Intelligent Data Analysis*. Springer, Berlin, Heidelberg, pp. 249–260.

Bottou, L. 2010 Large-scale machine learning with stochastic gradient descent. In: *Proceedings of COMPSTAT'2010* (Y. Lechevallier & G. Saporta, eds). Physica-Verlag HD, Heidelberg, Germany, pp. 177–186.

Candelieri, A. 2017 Clustering and support vector regression for water demand forecasting and anomaly detection. *Water* **9** (3), 224.

Chen, C., He, W., Zhou, H., Xue, Y. & Zhu, M. 2020 A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China. *Nature Scientific Reports* **10** (1), 3904.

Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J. R., Mellouli, S., Nahon, K., Pardo, T. A. & Scholl, H. J. 2012 Understanding smart cities: an integrative framework. In: *Proceedings of 45th Hawaii International Conference on System Sciences (HICCS 2012)*. IEEE Computer Society Press, Maui, HI, USA, pp. 2289–2297.

Di Nardo, A., Alcocer-Yamanaka, V. H., Altucci, C., Battaglia, R., Bernini, R., Bodini, S., Bortone, I., Bourguett-Ortiz, V. J., Cammissa, A., Capasso, S. & Cascetta, F. 2015a New perspectives for smart water network monitoring, partitioning and protection with innovative on-line measuring sensors. In: *Procceding of IAHR World Congress*, The Hague, The Netherlands.

Di Nardo, A., Di Natale, M., Gisonni, C. & Iervolino, M. 2015b A genetic algorithm for demand pattern and leakage estimation in a water distribution network. *Journal of Water Supply: Research and Technology – AQUA* **64** (1), 35–46.

Domingos, P. & Hulten, G. 2000 Mining high-speed data streams. In: *Proceedings of KDD 2000*. ACM, Boston, USA, pp. 71–80.

Friedman, J. H. 2002 Stochastic gradient boosting. *Computational Statistics & Data Analysis* **38** (4), 367–378.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M. & Bouchachia, A. 2014 A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)* **46** (4), 44.

García Valverde, D., Quevedo Casín, J. J., Puig Cayuela, V. & Saludes Closa, J. 2015 Water demand estimation and outlier detection from smart meter data using classification and Big Data methods. In: *2nd New Developments in IT & Water Conference*, Rotterdam, Holland, pp. 1–8.

Goodfellow, I., Bengio, Y. & Courville, A. 2016 *Deep Learning*. MIT Press, Cambridge, MA, USA.

Griffin, R. C. & Chang, C. 1990 Pretest analyses of water demand in thirty communities. *Water Resources Research* **26** (10), 2251–2255.

Hastie, T., Tibshirani, R. & Friedman, J. 2009 *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, USA.

Ikonomovska, E., Gama, J. & Džeroski, S. 2015 Online tree-based ensembles and option trees for regression on evolving data streams. *Neurocomputing* **150**, 458–470.

Ioannou, A. E., Kofinas, D., Spyropoulou, A. & Laspidou, C. 2017 Data mining for household water consumption analysis using self-organizing maps. *European Water* **58**, 443–448.

Jeong, J. & Park, E. 2019 Comparative applications of data-driven models representing water table fluctuations. *Journal of Hydrology* **572**, 261–273.

Kanakoudis, V., Tsitsifli, S., Papadopoulou, A., Curk, B. C. & Karleusa, B. 2016 Estimating the water resources vulnerability index in the Adriatic Sea region. *Procedia Engineering* **162** (Suppl. C), 476–485.

Kenda, K., Čerin, M., Bogataj, M., Senožetnik, M., Klemen, K., Pergar, P., Laspidou, C. & Mladenić, D. 2018a Groundwater modeling with machine learning techniques: Ljubljana polje aquifer. *Proceedings* **2** (11), 697.

Kenda, K., Koprivec, F. & Mladenić, D. 2018b Optimal missing value estimation algorithm for groundwater levels. *Proceedings* **2** (11), 698.

Kenda, K., Kažič, B., Novak, E. & Mladenić, D. 2019 Streaming data fusion for the Internet of Things. *Sensors* **19** (8), 1955.

Kofinas, D., Mellios, N., Papageorgiou, E. & Laspidou, C. 2014 Urban water demand forecasting for the island of Skiathos. *Procedia Engineering* **89**, 1023–1030.

Krause, J., Perer, A. & Bertini, E. 2014 INFUSE: interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics* **20** (12), 1614–1623.

Krause, J., Perer, A. & Ng, K. 2016 Interacting with predictions: visual inspection of black-box machine learning models. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, pp. 5686–5697.

Laspidou, C. 2014 ICT and stakeholder participation for improved urban water management in the cities of the future. *Water Utility Journal* **8**, 79–85.

Liu, H. & Motoda, H. 2007 *Computational Methods of Feature Selection*. CRC Press, Boca Raton, FL, USA.

Maier, H. R. & Dandy, G. C. 2000 Neural networks for the prediction and forecasting of water resources variables: a review of modeling issues and applications. *Environmental Modeling & Software* **15** (1), 101–124.

Manyika, J., Chui, M., Bisson, P., Woetzel, J., Dobbs, R., Bughin, J. & Aharon, D. 2015 *Unlocking the Potential of the Internet of Things*. McKinsey Global Institute, New York, NY, USA.

Mellios, N., Kofinas, D., Papageorgiou, E. & Laspidou, C. 2015 A multivariate analysis of the daily water demand of Skiathos Island, Greece, implementing the artificial neuro-fuzzy inference system (ANFIS). In: *E-Proceedings of the 36th IAHR World Congress*, The Hague, The Netherlands, pp. 1–8.

Nie, S., Bian, J., Wan, H., Sun, X. & Zhang, B. 2016 Simulation and uncertainty analysis for groundwater levels using radial basis function neural network and support vector machine models. *Journal of Water Supply: Research and Technology – AQUA* **66** (1), 15–24.

Oyebode, O. 2019 Evolutionary modeling of municipal water demand with multiple feature selection techniques. *Journal of Water Supply: Research and Technology – AQUA* **68** (4), 264–281.

Oza, N. C. 2005 Online bagging and boosting. 2005. In: *IEEE International Conference on Systems, Man, and Cybernetics*, Waikoloa, Hawaii, USA, pp. 2340–2345.

Recknagel, F. 2001 Applications of machine learning to ecological modeling. *Ecological Modeling* **146** (1–3), 303–310.

Rogers, P., De Silva, R. & Bhatia, R. 2002 Water is an economic good: how to use prices to promote equity, efficiency, and sustainability. *Water Policy* **4** (1), 1–17.

Sarle, W. S. 1994 Neural networks and statistical models. In: *Proceedings of the Nineteenth Annual SAS Users Group International Conference*. SAS Institute, Cary, NC, USA, pp. 1538–1550.

Soleimani, S., Bozorg-Haddad, O., Saadatpour, M. & Loáiciga, H. A. 2019 Simulating thermal stratification and modeling outlet water temperature in reservoirs with a data-mining method. *Journal of Water Supply: Research and Technology – AQUA* **68** (1), 7–19.

Swilling, M., Robinson, B., Marvin, S. & Hodson, M. 2013 *City-Level Decoupling: Urban Resource Flows and the Governance of Infrastructure Transitions*. A Report of the Working Group on Cities of the International Resource, UNEP (United Nations Environment Programme), Nairobi, Kenia.

Tiwari, M. K. & Adamowski, J. F. 2014 Medium-term urban water demand forecasting with limited data using an ensemble wavelet–bootstrap machine-learning approach. *Journal of Water Resources Planning and Management* **141** (2), 04014053.

Washburn, D., Sindhu, U., Balaouras, S., Dines, R. A., Hayes, N. M. & Nelson, L. E. 2010 *Helping CIOs Understand 'Smart City' Initiatives: Defining the Smart City, Its Drivers, and the Role of the CIO*. Forrester Research, Inc, Cambridge, MA, USA.

Zhang, Q., Yang, L. T., Chen, Z. & Li, P. 2018 A survey on deep learning for big data. *Information Fusion* **42**, 146–157.

First received 13 November 2019; accepted in revised form 25 March 2020. Available online 20 April 2020