Knowledge-Based Systems 222 (2021) 107012

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Hydro-power production capacity prediction based on machine learning regression techniques

C. Condemi [a,1], D. Casillas-Pérez [b,*,1], L. Mastroeni [a,1], S. Jiménez-Fernández [c,1], S. Salcedo-Sanz [c,1]

[a] Department of Economics, Roma Tre University, Rome, Italy
[b] Department of Signal Processing and Communications, Universidad Rey Juan Carlos, Fuenlabrada, Madrid, Spain
[c] Department of Signal Processing and Communications, Universidad de Alcalá, Madrid, Spain

ARTICLE INFO

Article history:
Received 20 July 2020
Received in revised form 26 March 2021
Accepted 30 March 2021
Available online 5 April 2021

Keywords:
Green power
Hydro-power capacity prediction
Machine learning
Artificial neural networks
Support vector regression
Feature reduction techniques
Energy markets

ABSTRACT

Hydro-power is a widespread source of energy, which currently provides over 60% of total renewable electricity production. As such, it plays a key role in green power generation, and has a fundamental influence on power market prices, because it can be used as a buffer for more volatile renewable sources, and it is relatively cheap to ramp up and down. For these reasons, it is of paramount importance to accurately predict the monthly hydro-power production capacity of wide geographical zones of the electricity market. In fact, future hydro-power production capacity depends on meteorological and climatic processes, water storage as result of pumping activity in the plant, and, of course, actual production, and this makes it extremely difficult to obtain an accurate prediction using traditional techniques, such as auto-regressive models. In this paper we propose a methodology based on machine learning (ML) regression techniques, mainly artificial neural networks and support vector machines, and feature reduction mechanisms, such as principal component analysis and feature grouping techniques. We apply these techniques to model the relationship between the meteorological and climatic variables and the total water in the reservoir used for the hydro-power generation. We show how ML regression techniques are able to obtain an accurate prediction of the hydro-power capacity in a real life example in Northern Italy.

1. Introduction

In an increasingly competitive market such as the electricity one, the ability to formulate bidding strategies plays a crucial role. At the same time, the transition to green electricity is hindered by the dependence on uncontrolled weather events (as shown in [1] for wind or in [2] for solar power), and the lack of storage technologies for electricity is a barrier to competitive bidding. In this regard, hydro-power storage plants play a key role within green power generation [3], because, in addition to being widely disseminated (they produce about 60% of renewable electricity generation [4]), by design they can accumulate the generating resource [5–7]. This helps mitigate the short-term productivity uncertainty related to the availability of renewable resources [8] and, moreover, allows for planning competitive strategies in the markets [9–11].

Hydro-power plants produce electricity by exploiting the power derived from the energy of falling or fast-running water, naturally influenced by rainfall and/or snowpack melting. Then, the sites with the highest concentration of such plants are the slopes and downstream areas of the mountain ranges. In fact, countries with the highest hydroelectric production are characterized by the presence of mountain ranges: Norway which, in 2020, generated 125 796 GWh from hydro-power sources, Brazil (397 877 GWh), Switzerland (40 962 GWh), Austria (44 223 GWh), Italy (47 499 GWh), Canada (381 908 GWh), Japan (88 005 GWh) and Russia (196 169 GWh) [4].

The bidding strategies of hydro-power producers represent one of the main drivers of hydro-power storage production [12–14]. These strategies are based on two main criteria: the first one, *profitability*, is based on today's production returns compared to future ones [14]. The second, *production capacity*, is based on the quantity of available water in the reservoir [9,12,14]. Since most of the storage hydro-power producers adopt these general criteria, the hydro-power supply by market zone is the result of a comparison between future and current scenarios of power market prices and production capacity [15]. Thus, the prediction of the production capacity by market zones is a key element to determine hydro-power supply and generation, and the latter is a key driver to predict power prices [16–18].

[*] Corresponding author.
E-mail address: david.casillas@urjc.es (D. Casillas-Pérez).
[1] All authors have contributed equally to the work.

https://doi.org/10.1016/j.knosys.2021.107012
0950-7051/© 2021 Elsevier B.V. All rights reserved.

Another important aspect associated with production capacity is that stored water resources are in some cases also used for activities other than electricity generation (such as agriculture, civil consumption, etc.). Thus, predicting hydro-power capacity on a monthly basis is a useful aid to policymakers who can have the information they need to act promptly, not only in terms of energy policies (i.e., to maintain the share of renewable energy and thus coordinate solar-wind-water strategies) but also in terms of policy strategies (i.e., strategic planning).

The problem of water availability prediction in reservoirs has been tackled with different computational methods: using time series processing algorithms [19], empirical orthogonal functions [20], error correction-based forecasting [21], multivariate approaches [22] or ensemble-based algorithms [23]. Machine Learning (ML) approaches have also been successfully applied to these problems. One of the first approaches using ML algorithms for water level prediction in reservoirs is [24], which compares the performance of artificial neural networks and neuro-fuzzy approaches in a problem of short-term water level in two German rivers, from hydrological upstream data. Adaptive neuro-fuzzy inference algorithms are also considered in [25] and in [26] for water level prediction in reservoirs after typhoons events. In [27] the performance of different ML algorithms, such as neural networks, support vector regression and deep learning algorithms, are evaluated in a problem of reservoir operation (mainly inflow and outflow prediction) at different time scales, in the Gezhouba Dam, across the Yangtze river, in China. In [28] a recurrent neural network is proposed for estimating the inflow of a reservoir from a distributed hydrological model. In [29] different ML regressors have been applied to a problem of water level prediction in a hydro-power plant in Galicia, Spain, using in-situ measurements of streamflow and precipitation. Several hybrid ML approaches have also been proposed for reservoir water availability issues, such as [30], where a neural network is hybridized with a genetic algorithm in order to improve the network training. This hybrid approach has been tested in a problem of dammed water level prediction at the low Han River, China. In [31] the authors propose a support vector regression (SVR) algorithm hybridized with a fruit fly algorithm for SVR parameters optimization. In this case, it has been tested in a problem of river flow forecasting in the lake Urmia basin, Iran.

Despite these important previous works on ML techniques applied to the prediction of water availability, there are very few works devoted to specific estimation of hydro-power production capacity. In particular, the majority of previous works discussing ML approaches to water availability deal with problems in hydrology rather than in renewable energy.

In general, the hydro-power production capacity of a geographical area is closely linked to the availability of water, which is in turn naturally affected by rainfall and/or snowpack melting. The latter is subsequently influenced by snowpack level, temperatures and solar radiation. The relationship between these predictors and the hydro-power capacity production is neither linear nor synchronous. This means that there is a time lag between when the meteorological event occurs and the subsequent update of water in the hydro reservoir [14]. In addition, there may be a persistence of the effect of the variables over time. These problems appear magnified when wide geographical areas of study are involved, such as different hydrological basins, for example. To this aim, it is important to analyze the geographical distribution of the plants, which is obviously not uniform, and then to select the predictors to be used according to the subareas involved.

In this paper, we present a study on hydro-power production capacity estimation with ML regression techniques, exemplified on Northern Italy, a wide power market zone. The total *production capacity* of a specific hydro-power plant is defined as the amount of electricity that could be produced, given the level of water in its reservoir. Inflows to reservoirs come from both natural contributions and pumping systems. In this paper we focus on modeling the production capacity from natural contributions. Different adaptations of the methodology are considered in order to improve the performance of the ML regression techniques in this problem. First, we consider meteorological input variables referring to each plant. The ML regressors can be launched directly with these variables as inputs, to obtain a first result. However, the number of predictive variables is in the order of hundreds. Furthermore, if several plants are present in the same hydrological basin, their variables are correlated, which can generate information redundancy and affect stability of the regressors. To avoid these problems, we grouped the variables by hydrological basin, reaching results that are more accurate than those obtained using the unstructured complete set of variables as direct inputs to the ML regressors. Furthermore, we also compare the results obtained by applying grouping with those by applying feature reduction methods. We show that, in the case study area, the proposed ML regression techniques, when applying grouping, provide a good modeling and estimation of the hydro-power production capacity, within a monthly prediction time-step.

The remainder of the paper has been structured in the following way: Section 2 details the definition of the problem. Section 3 provides a summary the ML methods applied to this problem of hydro-power production capacity prediction. Section 4 presents the results of fitting the algorithms for the Northern Italy case study. Section 5 is devoted to the discussion of the results while Section 6 closes the paper with some final conclusions and remarks on the research carried out. Appendices A and B show a summary of the acronyms and metrics used along the paper. Finally, Appendix C includes some additional data for the experiments carried out.

2. Problem statement

A crucial phenomenon for water availability in hydro-power systems is represented by snowpack melting, which is strongly dependent on micro-area conditions (snow texture, exposure of the area to sunlight, soil morphology, wind, etc.). However, considering all these micro-territorial specifications, the model is complex and not easy to generalize. For this reason, the meteorological variables considered in this work are those influencing the phenomenon from a macro-territorial point of view, namely rainfall, solar radiation, temperature and snowpack level. This approach makes the model practical and easy to generalize to similar geographical zones.

2.1. Problem definition

The aggregate production capacity S^T of all plants located in a wide geographical area can be expressed by the following linear flow balance equation:

$$S^T = S^N + S^P - SG, \qquad (1)$$

where S^N represents the production capacity due to the water availability from natural contributions, S^P is the production capacity due to artificial water availability (i.e., resulting from pumping activity), and SG represents the total amount of electricity generated by storage hydro-power plants in the whole area of study, during the period examined. The production capacity decreases when plants produce electricity SG, while it increases when water flows into the reservoirs. In particular, the one resulting from natural contributions S^N is affected by the meteorological variables that influence the water cycle, that is, rainfall and snowpack melting.

As already mentioned, the variables considered in this work are those influencing the phenomenon from a macro-territorial point of view. For these reasons, we can express the aggregate production capacity from natural contributions S^N as a general non-linear function of rainfall Rn, snowpack SW, temperature T and solar radiation IR as follows:

$$S^N = f(Rn, SW, T, IR) \tag{2}$$

Thus, the objective of this work is to estimate S^N by modeling the function f using ML regression techniques (i.e., using measurements of rainfall Rn, snowpack SW, temperature T and solar radiation IR as input variables). Obviously, in other areas of study, alternative variables such as evapotranspiration could be a key factor to be taken into account. In the present case, however, rainfall, snowpack, temperature and solar radiation are enough to obtain an accurate modeling of the hydro-power capacity.

It is important to remark, however, that there is not a direct correspondence between S^N and the weather phenomena affecting precipitation and snowpack melting, such as, Rn, SW, T or IR. In fact, the inflow that such phenomena generate incurs in losses, before turning into hydroelectric productivity. First of all, the inflows generated by rainfall and snowpack melting suffer losses on their way to the reservoirs. In particular, part of the inflow is absorbed by the soil. In addition, not all the water that flows into the reservoirs participates in the process of power production (e.g. due to pipe leaks, evaporation). Thus, the water volume does not correspond directly to the available production. This is why it is necessary to estimate the capacity production on the basis of available data and actual production.

Performing regression of the unknown function f needs input–output pairs of data for a training step. According to Eq. (1), S^N can also be expressed as a function of the total production capacity S^T, the capacity of the pumping water S^P and the hydroelectricity generation of the plants SG. Consequently, we can use this equation to obtain the ground truth for S^N and train the supervised ML regressors to model function f of natural hydro-power production capacity. Once we have this modeling function f, we are able to estimate the natural production capacity of the system within a given time horizon.

Regarding the considered variables affecting the transformation of snow into water resources, in this study we address local variables, specific to each hydro-power plant, such as snowpack SW, solar radiation IR and temperature T. In this way, we select the variables to represent only the information related to the areas where the hydro-power plants are located. Rainfall, differently from snowpack, is distributed in a very unpredictable way. Consequently, the data recorded by each rain gauge have to be processed to obtain an estimate of the rainfall flowing into the streams associated with each reservoir. For this reason, considering rainfall estimations that affect each reservoir makes the problem very complex. In this study we propose to consider separately the estimation of rainfall collected in the mountain ranges, Rn^A, and those collected in the rest of macro-area considered, Rn^N.

We set the time step and the prediction horizon to one month, since most variables, including the target S^N, are measured or obtained at this temporal resolution. This resolution is enough to study the effects of predictive variables in hydro-power capacity production. Thus, having specified the time horizon, Eq. (2) can be written as the following:

$$S^N_m = S^T_m - S^P_m + SG_m \tag{3}$$

Consequently, the problem we address in this paper is the modeling of the function $S^N_m = f(\mathbf{Rn}_m, \mathbf{SW}_m, \mathbf{T}_m, \mathbf{IR}_m)$, that enables the estimation of the monthly natural hydro-power production capacity over a wide geographical area with the dependencies defined as:

$$\begin{cases} \mathbf{Rn}_m = (Rn^A_m, Rn^N_m) \\ \mathbf{SW}_m = (SW^{(1)}_m, \dots, SW^{(i)}_m, \dots, SW^{(n)}_m) \\ \mathbf{T}_m = (T^{(1)}_m, \dots, T^{(i)}_m, \dots, T^{(n)}_m) \\ \mathbf{IR}_m = (IR^{(1)}_m, \dots, IR^{(i)}_m, \dots, IR^{(n)}_m). \end{cases} \tag{4}$$

The index n denotes the number of hydro-power plants and m the calendar month.

3. Methods considered

We compare the performance of multiple ML algorithms on a specific case study in Northern Italy. We have considered multi-layer perceptron (MLP) networks [32,33], extreme learning machines (ELM) [34] and support vector regression algorithms (SVR) [35,36]. In addition, we have also implemented a linear regressor (REG) as baseline for comparison.

We selected these algorithms because they present a trade-off between modeling complex nonlinear input–output relations and using a fair number of parameters to train. As we will see in Section 4.3, the case study relies on a reduced amount of data. For this reason, we discarded Deep Neural Networks (DNN), such as Recurrent Neural Network (RNN), Convolution Neural Network (CNN), or even the new Long-Short Term-Memory (LSTM) or Gated Recurrent Unit (GRU), which demand a large amount of data. The same issue occurs with boosted tree models, such as random forest or decision trees. ARIMA models were also discarded, since they model linear system responses and do not present good results with dependent variables, not even when considering ARX to introduce exogenous variables.

All the tested methods are supervised ML regressors, trained by a labeled set of input–output pairs of data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. The output of the trained ML models is the available production capacity for the next month (from Eq. (3)):

$$y_i = S^N_{m+1} = f(\mathbf{x}_i) = f(\mathbf{Rn}_m, \mathbf{SW}_m, \mathbf{T}_m, \mathbf{IR}_m), \quad 1 \le i \le n \tag{5}$$

where n represents the total number of months in the database. Let us now briefly summarize the most important characteristics of the different prediction algorithms considered in this paper, and also the one used for feature selection.

A multi-layer perceptron is a type of feed-forward artificial neural network (ANN) that has been successfully applied to non-linear classification and regression problems [32,33]. An ANN implements a set of neurons, nodes of a graph, fully interconnected with each other through synaptic connections with different assigned weights. These nodes are organized in three main type of layers: the input layer, some hidden layers connected in cascade, and an output layer. During the training phase, the training algorithm assigns values to these weights to minimize the error between the estimated output $\tilde{y}$ and the expected output y. For this process there are different types of iterative training algorithms, such as the *backpropagation* or the Levenberg–Marquardt algorithms [37].

The extreme-learning machine (ELM) [34] is a fast training method mainly used for feed-forward multi-layer perceptron structures formed by a single hidden layer just like MLPs (see [34,38] for details on the algorithm). The ELM algorithm's efficiency is close to the linear regressors in terms of computation time. However, its capacity of modeling complex functions is somehow reduced compared with the MLPs. Note that reducing internal degrees of freedom is usually an advantage for small data sets.

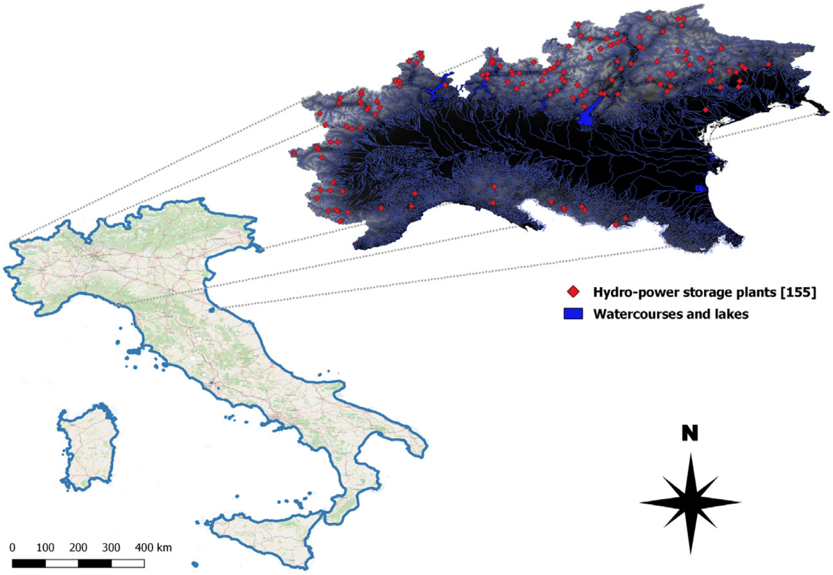


Fig. 1. Distribution of the hydro-power plants and hydrography of Northern Italy.

The support vector regressor (SVR) is one of the most popular ML methods used for regression, curve fitting or even more general function approximation problems [35,36]. It is based on the well-known support vector machine methodology. We also include it among the ML regression algorithms tested, since it has obtained excellent results in different regression problems with small data sets. The SVR algorithm has only 3 hyper-parameters to be tuned for its application to regression problems [35].

The linear regressor (REG) [39] assumes that there exists a linear relationship between the dependent variable y and the explanatory variables (the independent variables) $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^N$. It can be expressed by the formula: $y = \beta_0 + \mathbf{x}^T \boldsymbol{\beta}$. This method is attractive because there is a direct relationship between coefficients and feature importance, and it is more interpretable than other models. The linear regressor retrieves its coefficients minimizing the mean squared error of the residual, which yields a closed-form solution: $\beta = (X^T X)^{-1} X^T Y$.

We specifically use the implementations provided by MATLAB language program (version R2019b with the *Statistics and Machine Learning Toolbox*) for the SVR, REG and MLP methods. The code implementation for the ELM is based on [40].

Another important method considered is the so called Principal Components Analysis (PCA) [41], which allows significantly reducing the number of input variables in our prediction problem. PCA is strongly linked to the singular value decomposition of the input matrix. This technique looks for the linear subspace which best represents the variability of the inputs using less features. It computes the singular value decomposition and orders the singular values in descending order $\{\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots \sigma_n = 0\}$. Then, it is possible to discard the less important values based on PCA results, retaining the best subspace representation. A threshold for this information removal must be set by the user.

4. Experiments and results

In this section we show the results obtained by different ML regression techniques in the case study of monthly hydro-power capacity estimation in Northern Italy. First, we describe the database available for Northern Italy in Section 4.1, a grouping procedure to reduce the number of input variables of the problem in Section 4.2 and the considered experimental setting in Section 4.3. Then, we show the results of the correlation study among the input variables in Section 4.4. Finally, we detail the results obtained by the ML regression techniques tested in Section 4.5 and Section 4.6, distinguishing between using all input variables or grouped input variables, respectively.

4.1. Database description: a case study in Northern Italy

In this paper we consider as case study the hydro-power plants of the *power market zone* of Northern Italy,[2] see Fig. 1.

In this market zone (corresponding to 96 392.17 km^2 [42]), we identify and geo-reference 155 hydro-power storage plants (see Fig. 1). The total installed capacity of these plants is 12.40 TWh. However, this power is not evenly distributed neither among the individual plants nor over the entire geographical area. In fact, according to Table 1, we observe that up to 96.11% of the installed capacity is located in the geographical areas of the Alps-Italy, which represents the mountain range of the Northern Italy region. Therefore, it is clear that the plants taken into account are mainly influenced by mountain meteorological and climatic processes.

Table 1 and Fig. 2 show the strong variability of the individual installed capacity of the considered hydro-power plants. The 69.56% of the total installed capacity is associated with large plants (plants with a installed capacity larger than 100 MW), whereas medium plants (installed capacity between 10 and 100 MW) and small plants (installed capacity lower than 10 MW) represent 30.43% and 0.012%, respectively. This classification is of paramount importance for the analysis of production capacity, since, generally, the size of the reservoirs is proportional to the plants' installed capacity.

As presented in Section 2.1, for each plant we construct an indicator of the monthly mean of the snowpack depth (SW^i_m), of the monthly global solar radiation (IR^i_m) and of the monthly mean of the temperature (T^i_m). These values depict the average level of the variables considered, along the watercourses that feed the reservoir of the hydro-power plant. In addition, in order to

[2] This power market zone is the geographic area composed of the following regions: Valle d'Aosta, Piemonte, Liguria, Trentino, Veneto, Friuli Venezia Giulia and Emilia Romagna.

Table 1
Hydro-power plants characteristics in Northern Italy.

Size	N. plants	Installed capacity				β
		Total (MW)	Alps-Italy (%)	Other (%)	Mean (MW)	Std. Dev. (MW)
Large plants ($\beta > 100$ MW)	30	8628.31	96.74	3.26	287.61	243.09
Medium plants ($10 < \beta \leq 100$ MW)	110	3638.08	96.18	3.82	33.07	21.00
Small plants ($\beta \leq 10$ MW)	15	138.48	54.59	45.41	9.23	0.76
Total	155	12 404.87	96.11	3.89	80.03	148.80

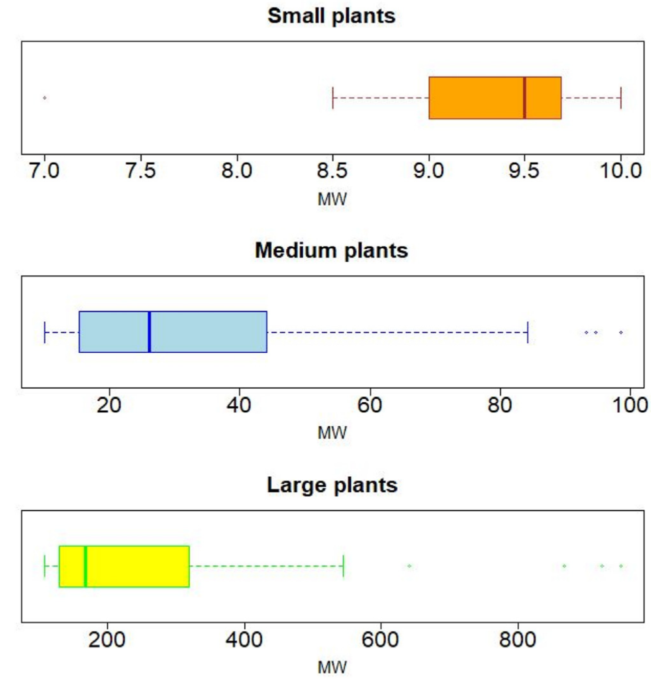


Fig. 2. Installed capacity box-plot.

represent the rainfall in the area under consideration, we have divided the study area into two sub-areas: Alps-Italy and the rest of Northern Italy. Then, for each of the two sub-areas, we have considered the estimated monthly rainfall Rn_m^A and Rn_m^N, respectively.

Moreover, according to Eq. (3), the ground truth is computed from the monthly values of hydro-power production capacity for almost 6 years (64 months continuously acquired from September 2013 to December 2018).

Thus, specifically, in this case study:

- S_m^T is the total of monthly production capacity of all hydro-power storage plants operating in Northern Italy, i.e., the hydroelectricity that can be produced with the water available in the reservoir during the mth month.
- S_m^P represents the total of hydroelectricity that can be produced with the water obtained through pumping activities during the mth month, in Northern Italy. It can be estimated by multiplying monthly pumping electricity consumption in Northern Italy (P_m) by a conversion coefficient (η) estimated from year to year as the ratio of the annual generation of pure pumped-storage plants to their annual electricity consumption.
- SG_m represents the total amount of electricity generated by storage hydro-power plants in Northern Italy, during the mth month.

4.2. Grouping the input variables

The meteorological variables considered in Section 4.1 amount to a total of 467 (there are 155 plants with 3 variables per plant plus 2 global variables for the rain). This huge number of variables managed by ML regression techniques possibly leads to over-fitting or over-training problems. Moreover, plants in a hydrological sub-basin are quite close to each other. Therefore, these plants often draw their resources from the same watercourses and the meteorological indices considered are very similar. This implies a high redundancy in the information contained in the series. In order to reduce the number of variables without loss of useful information, we have grouped the variables related to snow-melting into 28 groups, considering hydrographic sub-basins. Fig. 3 shows the cartography of the sub-hydrographic basins in Northern Italy and the related group subdivision.

Each group contains the hydro-power plants belonging to a different hydrological sub-basin, except for three cases in which there is a high spatial contiguity among the sub-basins. In these cases, we have merged into the same group the plants located in two neighboring sub-basins, obtaining the following groups: TS, CX (both to the South West) and Y (to the North East).

Thus, for each group we have considered the average values of SW_m^i, T_m^i and IR_m^i, with $i = 1, \dots, n^*$, where n^* stands for the number of plants in a given group of the 28 considered.

4.3. Experimental setup

We have a monthly database (64 months) of the input variables involved in the problem and in the objective ground truth, continuously acquired from September 2013 to December 2018 (see Section 4.1 for a detailed explanation of the dataset). We have split the database into test and train sets, with a fraction $\alpha = 20\%$. The first 52 months (80% of the dataset) are used to train the proposed ML algorithms, and the last 12 months (20% of the dataset), which correspond to year 2018, are used to test the models. The statistical distributions of the test and train sets are quite similar.

We applied the SVR algorithm with Gaussian, linear and polynomial kernels (referenced as SVRg, SVRl and SVRp, respectively). In all these cases, the training algorithm used a K-fold cross-validation $K = 5$ to select the SVR hyper-parameters. We also considered the MLP and, due to the small amount of data, we have used a structure of two single hidden layers, with 10 neurons each after a scanning process between 5 and 25, with a K-fold cross-validation $K = 5$. Instead, in the ELM (which considers a perceptron-like structure with a single hidden layer) we have estimated the best number of neurons in the hidden layer which results in 19. Finally, conventional linear regressors have been applied, in order to have a baseline comparison technique, named as REG hereafter.

We have divided the experiments into two sets: training the ML algorithms with *all variables* as inputs, or using only the mean value computed in each *group of variables*, see Sections 4.1 and 4.2, respectively. Then, we have established the following four different experimental settings:

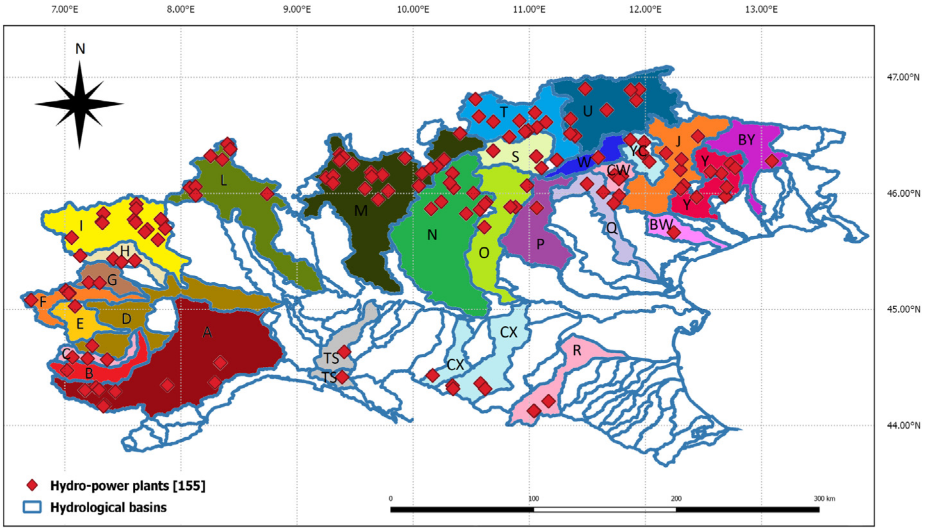
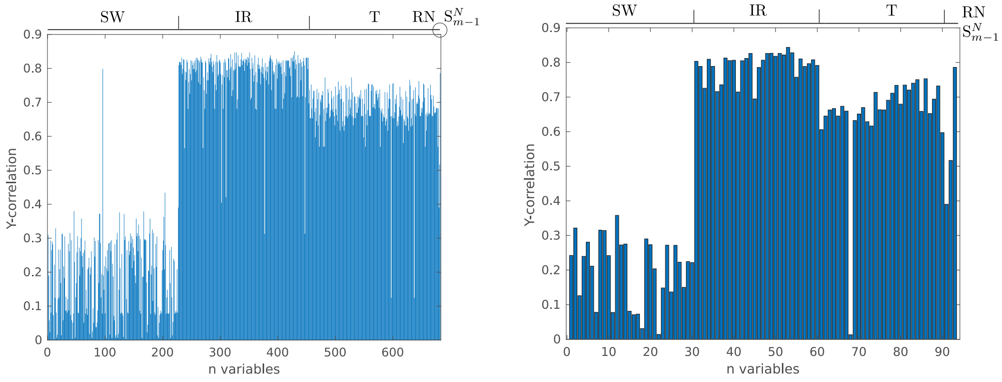


Fig. 3. Northern Italy sub-hydrographic basins and the grouping of hydro-power plants. Geographical coordinate system: WGS 84.



(a) All input variables vs output variable.

(b) Grouped input variables vs output variable.

Fig. 4. Correlation of the input variables versus the output variable, considering (a) all input variables, (b) grouped input variables. Last three columns show the correlation between the *RN* variables and S^N_{m-1} with the output S^N_m.

1. Hydro-power capacity prediction using all variables.
2. Hydro-power capacity prediction with the grouped variables.
3. Reducing the number of variables using PCA for the case of all variables.
4. Reducing the number of variables using PCA for the case of grouped variables.

To evaluate the quality of the obtained results, we provided the most common metrics used in ML studies: the root mean square error (RMSE), the mean absolute error (MAE) and the Newman–Pearson coefficient, also known as correlation coefficient (CORR). Predictions whose MAE and MSE are below 0.4 TWh can be considered as very good. Above 1 TWh are considered very poor predictions. A CORR value close to 1 corresponds to a very good prediction and a value below 0.6 is considered a poor one. Note that the evaluated methods penalize outliers and large deviations differently. A thorough description of these metrics can be found in [43], while their explicit formulation are presented in Appendix B.

4.4. Correlation structure among the variables of the problem

Any supervised ML method requires that the output variable to be predicted and the input variables used to infer it, are dependent. Thus, we have measured such a dependence relationship between the output and the input variables by computing their correlation. Fig. 4 shows these results for both sets analyzed: (a) using all the variables and (b) using the grouped variables.

In both cases, it can be seen that most variables are highly correlated. In particular, the highest correlation appears in rainfall *RN*, solar radiation *IR* and temperature *T* variables (around 0.5, 0.8 and 0.7 respectively), while snowpack depth *SW* shows less correlation (around 0.25) with the output variable. The correlation with temperature for group D is much lower than the rest. In this group the installed capacity of the plants is low (25 MW), so these plants contribute marginally to variations in zonal total production capacity. However, we keep it for consistency. A specific statistical hypothesis testing which evaluates whether the correlation is significantly different from zero can be found in Appendix D.

Nevertheless, it is known that large correlation values among features may affect stability of both linear and non-linear regressors, which suggests that a Principal Component Analysis (PCA) could help improve their performance. Therefore, we have computed the Pearson correlation matrix among all the different input variables, shown in Fig. 5. Correlation close to 1 is shown using white color, while total uncorrelation is shown using black color. Therefore, white stripes prove the existence of highly correlated variables. On the contrary, black stripes correspond to highly uncorrelated variables.

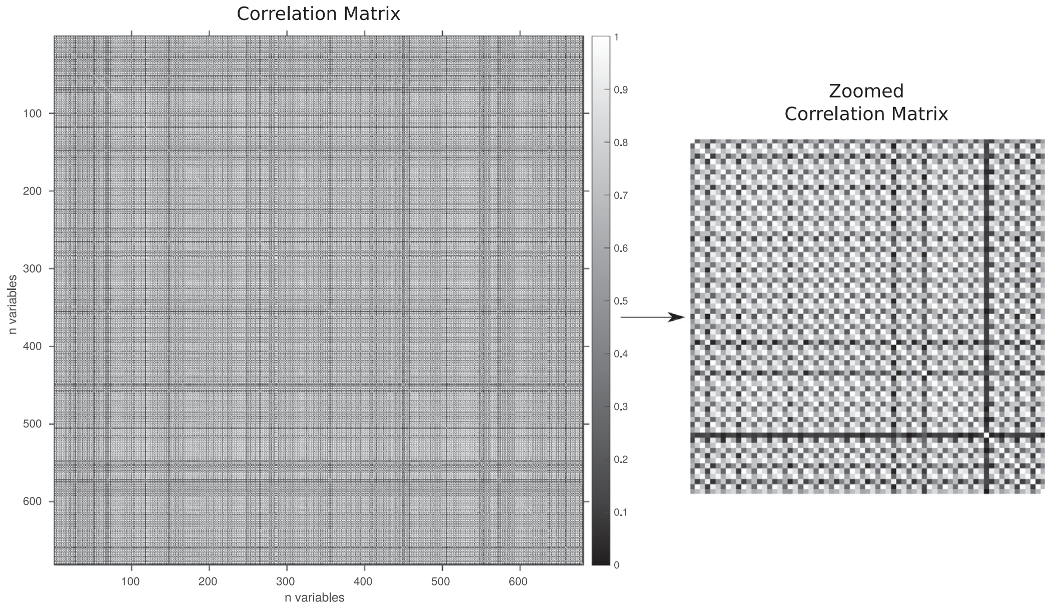


Fig. 5. Correlation matrix of all variables involved in the problem. Correlations close to 1 are shown using white color, whereas total uncorrelation is shown using black color. Therefore, white stripes prove the existence of highly correlated variables. On the contrary, black stripes correspond to highly uncorrelated variables.

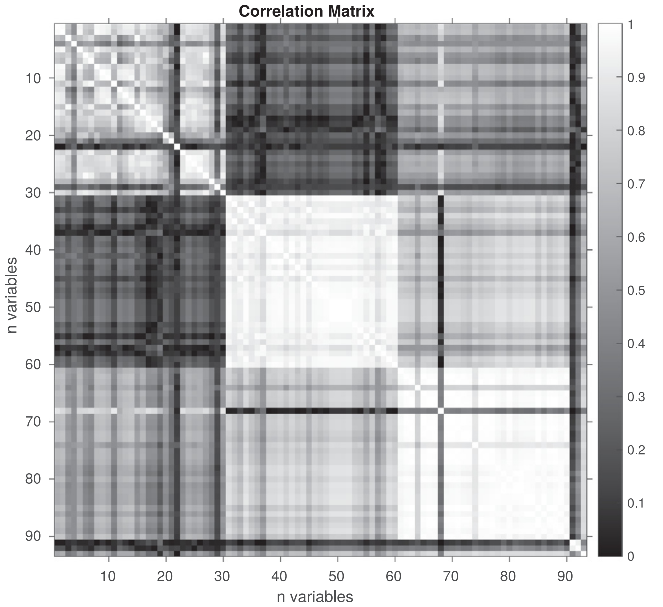


Fig. 6. Correlation matrix of grouped variables (93 variables). White straight lines prove the existence of highly correlated variables. On the contrary, black lines correspond to highly uncorrelated variables with the others.

A similar analysis, but using the grouped variables, is presented in Fig. 6 (93 input variables per month). There also exists some correlation among the grouped variables of the problem, but to a lesser extent than in the previous case. This can be due to the fact that nearby groups share exogenous variables, such as solar radiation, temperature or snowpack, as groups of variables which are far from each other are less correlated.

So far, we stated that several input variables used for training the ML methods are strongly correlated, even using the grouped variables. The main objective of PCA is to find a subspace that contains most of the information relevant to the problem. We have to establish a trade-off between the number of principal components used as inputs for the ML methods and the reduction of the number of variables. A high number of principal components results in a small reduction of the number of variables and in an inefficient removal of correlated variables. On the other hand, a reduced number of principal components may lead to a reduced prediction accuracy.

We then applied a PCA analysis to the variables, both using all variables and using only grouped variables. In both sets, we varied the number of principal components from 2 to 7, using the corresponding components to train ML methods.

4.5. Study considering all the input variables

Fig. 7 shows the prediction's performance of the proposed ML methods using all the input variables considered (numerical results are given in Table C.6 in Appendix C), as well as the ground truth (GT). SVR and REG are deterministic methods, i.e., they always obtain the same results when they are run with a given training and test set. However, both ELM and MLP are stochastic approaches, i.e., they can obtain slightly different results in different runs of the algorithms, as they are trained by iterative algorithms with random initialization. As a consequence, the ELM and MLP results shown in Fig. 7 are one of their representative realizations: in this case we have chosen the closest to their corresponding mean error. As can be seen in this figure, the evaluated models are able to capture input–output relationships in a training dataset. However, test results are generally not good for any of the regressors tested. These results indicate some sort of overfitting in the training process, maybe caused by the lack of input data compared to the number of features involved in the process, which is a known issue [44]. Therefore, predictions obtained could be improved.

Using a PCA analysis (stated in Section 4.4) and varying the PCA coefficients in a range from 2 to 7, we have trained again the ML methods, obtaining the numerical values shown in Table 2. Generally speaking, the use of PCA allows obtaining significant (useful) results. Specifically, the RMSE and MAE are now around the 0.7 and 0.6 TWh respectively, and the Pearson correlation is around 0.8. The best results have been obtained using 2 and 3 PCA components for all regressors studied.

When using 2 PCA components, the SVR with a linear kernel (SVRl) achieves the best results among all the regressors studied (see Table 2).

Table 2
CORR, RMSE and MAE results obtained using PCA on all input variables and varying the number of components between 2 and 7.

PCA	Method	SVRg	SVRl	SVRp	ELM	REG	MLP
2	CORR	0.7754	**0.8433**	0.7669	0.6417 ± 0.1800	0.8339	0.6003 ± 0.0691
	RMSE	0.8761	**0.7214**	0.8585	1.1231 ± 0.37248	0.7425	1.2187 ± 0.2699
	MAE	0.7247	**0.5992**	0.7211	0.8963 ± 0.2418	0.6110	1.0083 ± 0.0996
3	CORR	**0.8384**	0.8341	0.7608	0.6164 ± 0.1574	0.8333	0.7887 ± 0.0203
	RMSE	0.8358	0.7613	0.9044	1.1911 ± 0.27682	**0.7431**	0.8588 ± 0.2824
	MAE	0.7052	0.6233	0.7678	0.9683 ± 0.2067	**0.6121**	0.6983 ± 0.0379
4	CORR	0.8134	0.7535	**0.8265**	0.7331 ± 0.1383	0.7525	0.8097 ± 0.0724
	RMSE	0.9504	0.9707	0.9374	1.0417 ± 0.21740	**0.9194**	0.9473 ± 0.3037
	MAE	0.7896	0.7891	0.7579	0.8269 ± 0.1503	**0.7295**	0.7758 ± 0.0701
5	CORR	0.6301	0.7327	0.6973	0.6107 ± 0.1907	**0.7364**	−0.158 ± 0.2314
	RMSE	1.0869	1.0365	1.0025	1.1722 ± 0.29226	**0.9685**	2.0205 ± 0.4271
	MAE	0.8471	0.7590	0.8229	0.8866 ± 0.1890	**0.6919**	1.4331 ± 0.2397
6	CORR	0.5843	0.7346	0.7331	0.5841 ± 0.2305	**0.7411**	0.4513 ± 0.1277
	RMSE	1.0755	1.0318	1.0605	1.3803 ± 0.46418	**0.9518**	1.4766 ± 0.3338
	MAE	0.8567	0.7584	0.7081	1.0196 ± 0.2993	**0.6895**	0.9896 ± 0.1183
7	CORR	0.5560	0.6733	0.7053	0.4839 ± 0.2500	**0.7214**	0.6105 ± 0.1362
	RMSE	1.0904	0.9738	1.0153	1.8238 ± 0.85917	**0.9222**	1.5399 ± 0.4363
	MAE	0.8159	0.7670	**0.7000**	1.2644 ± 0.5376	0.7393	1.0595 ± 0.3388

Table 3
Results of the considered ML regression techniques evaluated using the grouped variables.

Method	CORR	RMSE	MAE
SVRg	0.6886	1.0982	0.9028
ELM	0.6983 ± 0.1992	1.2165 ± 0.3572	0.9519 ± 0.2822
REG	0.3803	4.234	2.9688
MLP	**0.9735 ± 0.0164**	**0.2593 ± 0.0315**	**0.2128 ± 0.0275**

On the contrary, MLP and ELM get the worst results among all the regressors tested. Surprisingly, the ELM gets even worse results than using all the variables as inputs. Using 3 coefficients, the SVRg with a Gaussian kernel obtains the best correlation result, 0.8384. Fig. 8 shows the prediction of the proposed ML methods using 2 coefficients PCA analysis on all the variables.

The reduction of the variables brings significant lower errors, and leads to accurate regressors, since it removes the overfitting problem. However, although the fits have improved (Fig. 8), the RMSE and MAE still do not indicate very good fits, so we further proceed with a grouping of the variables in order to improve the prediction.

4.6. Study considering the grouped variables

As described in Section 4.2, grouping the plants reduces the amount of input variables, therefore, we have considered 28 different groups of plants. For each group, we have averaged the temperature T_m, the solar radiation IR_m and the amount of snowpack SW_m. Moreover, we have added the rainfall on Alps-Italy Rn^A_m, the rainfall on the rest of Northern Italy Rn^N_m and the production capacity registered in the previous month S^N_{m-1}. Thus, the total number of input variables per month is now 93.

Fig. 9 shows the prediction of the proposed ML methods using the grouped variables as inputs. One can observe improved predictions for the training data. Predictions of never observed months (test dataset) are also accurate when considering the MLP approach.

Table 3 shows the metrics for the test set. Using the variables of the different groups, the best results in terms of correlation, RMSE and MAE are obtained by the MLP with 0.9735, 0.2593 TWh and 0.2128 TWh, respectively, which represent a good capacity prediction, the best achieved in this paper.

Reducing the correlation among the input variables significantly improved the MLP in the ungrouped case. Thus, we performed the same procedure for the grouped case, and applied the

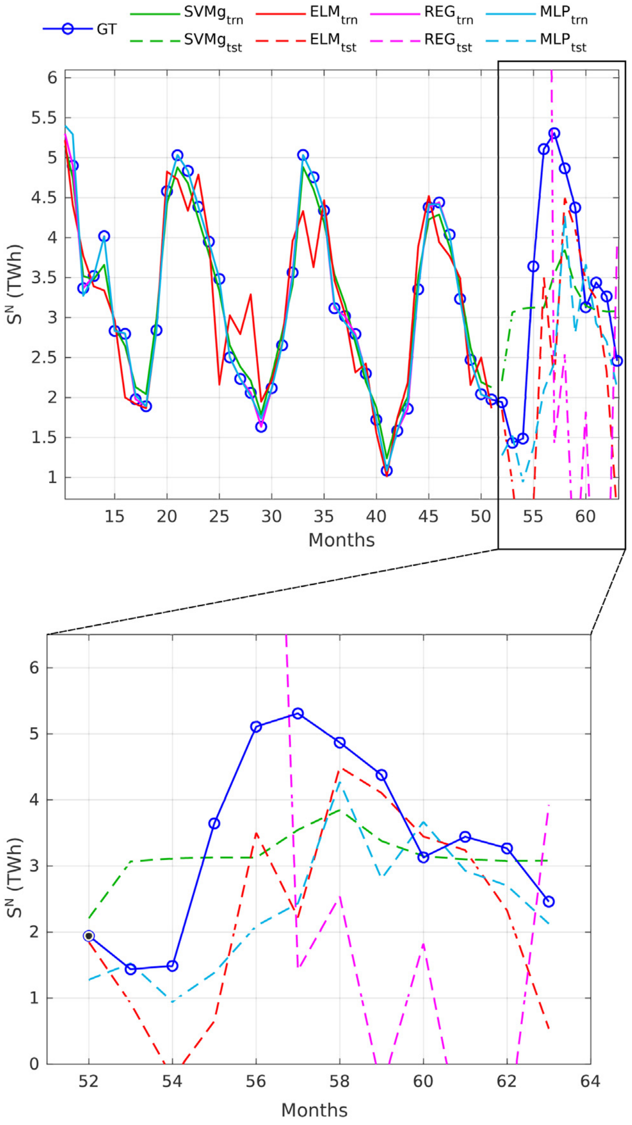


Fig. 7. Results of the considered regression techniques using all the input variables considered. Last twelve months are used to test the ML regression techniques. GT stands for ground truth for comparison purposes.

Table 4
CORR, RMSE and MAE results obtained by the evaluated ML methods varying the PCA components between 2 and 7, in case of using grouped variables.

PCA	Method	SVRg	SVRl	SVRp	ELM	REG	MLP
2	CORR	0.8289	0.8141	**0.8406**	0.4440 ± 0.2363	0.8184	0.8093 ± 0.0169
	RMSE	0.7374	0.7794	**0.7146**	2.7296 ± 2.8564	0.7626	0.7880 ± 0.2251
	MAE	0.6112	0.6497	**0.5965**	1.4428 ± 0.9497	0.6409	0.6544 ± 0.0235
3	CORR	**0.8087**	0.8000	0.7996	0.6890 ± 0.1704	0.7971	0.7950 ± 0.0221
	RMSE	**0.7978**	0.8204	0.8099	0.9755 ± 0.2796	0.8202	0.8196 ± 0.2841
	MAE	**0.6390**	0.6577	0.6563	0.7346 ± 0.1243	0.6566	0.6559 ± 0.0254
4	CORR	0.6838	0.7719	0.5832	0.3394 ± 0.3627	**0.7919**	0.5913 ± 0.0934
	RMSE	0.9643	0.9341	1.2097	1.6546 ± 0.6642	**0.8457**	1.0652 ± 0.3600
	MAE	0.7605	0.6637	0.8176	1.1021 ± 0.3539	**0.6341**	0.8620 ± 0.0675
5	CORR	0.6107	0.7740	0.7417	0.5361 ± 0.2690	**0.7926**	0.4454 ± 0.0538
	RMSE	1.0376	0.9342	0.9408	1.3522 ± 0.5977	**0.8449**	1.2965 ± 0.3034
	MAE	0.8466	0.6834	0.7695	0.9549 ± 0.3229	**0.6335**	0.9061 ± 0.0490
6	CORR	0.6305	0.8367	0.6979	0.5709 ± 0.2815	**0.8371**	0.7351 ± 0.0389
	RMSE	1.0178	0.8121	1.0085	1.2459 ± 0.5266	**0.7794**	1.0367 ± 0.3407
	MAE	0.8246	0.6553	0.7877	0.9211 ± 0.3278	**0.6037**	0.7895 ± 0.0489
7	CORR	0.6629	**0.8240**	0.6025	0.5836 ± 0.2408	0.8170	0.3786 ± 0.1012
	RMSE	0.9860	0.8357	1.4294	1.1996 ± 0.3947	**0.7910**	1.8218 ± 0.5179
	MAE	0.7634	0.6654	1.1029	0.8761 ± 0.2678	**0.6046**	1.2377 ± 0.2347

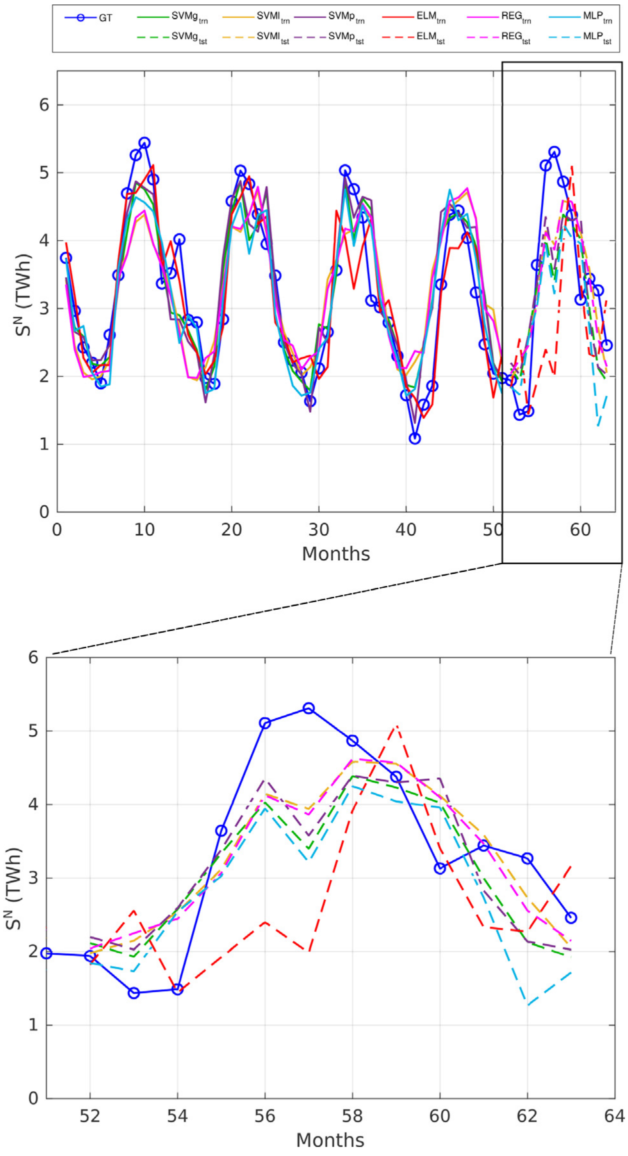


Fig. 8. Results of the ML regression techniques setting PCA components to 2 and using all the variables.

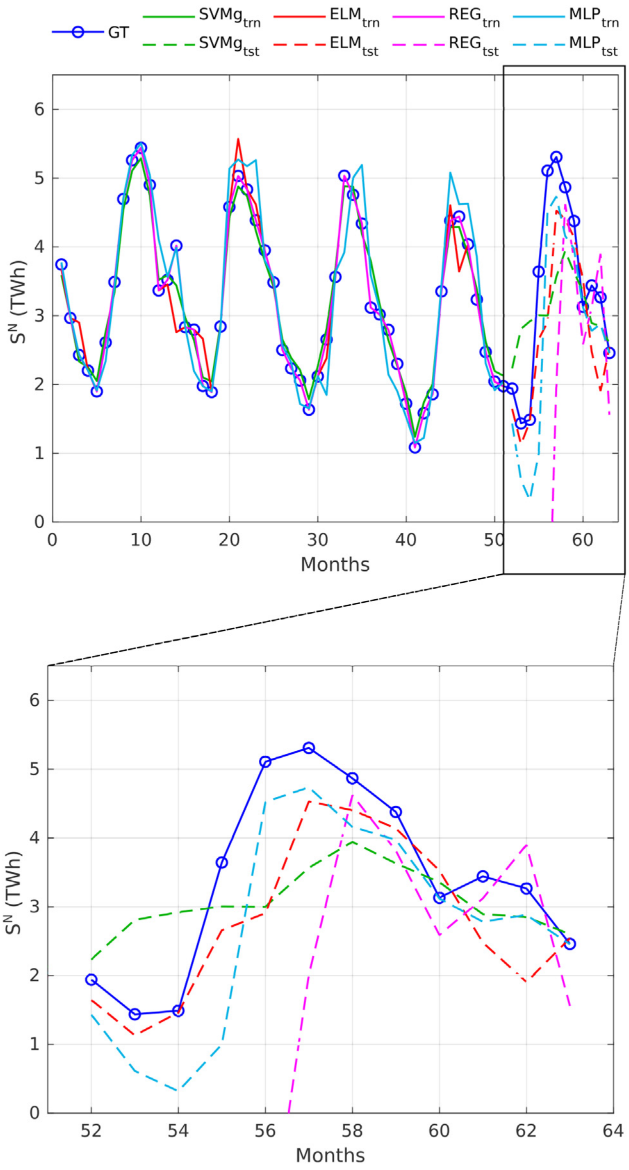


Fig. 9. Train and Test results by considering grouping of input variables.

PCA analysis varying the number of components between 2 and 7. The correlation, RMSE and MAE values obtained in this case are shown in Table 4.

As in the PCA analysis with all the variables, we observe that the best results in terms of correlation, RMSE and MAE are achieved with a low number of PCA coefficients, i.e., with 2 or 3 coefficients. Unlike the previous case (when all variables were considered), the reduction of the number of variables does not lead to an improvement of the results, compared to the use of all the grouped variables. In fact, the MLP obtains worse results with the PCA components as input variables than considering all the variables as inputs. Using 2 PCA coefficients, the SVR with polynomial kernel (SVRp) leads to the best results in terms of correlation, RMSE and MAE, with values of 0.8406, 0.7146 TWh and 0.5965 TWh, respectively. The SCRg, SVRl, MLP and REG report good results, close to the SVRp. In this case, the ELM obtains the worst results. In the case of 3 PCA coefficients, the SVR with Gaussian kernel (SVRg) is the one which leads to the best results in terms of correlation, RMSE and MAE, with 0.8087, 0.7978 TWh and 0.6390 TWh, respectively. The SVRp, SVRl, MLP and REG report good results, close to the SVRp. The ELM improves its results compared to the use of 2 coefficients, but it still obtains the worst results among all the regressors tested. Using more coefficients does not lead to better results, as can be seen in Table 4. Fig. 10 shows the prediction of the proposed ML methods using 2 PCA coefficients with the grouped variables.

The best regressor performance (MLP) is obtained without feature reduction PCA on the grouped variables. A possible explanation for this is that the PCA removes part of the non-linear relationships among the variables, which seem to be efficiently processed by the MLP in order to finally obtain good results for the problems. However, these relationships are not found by the MLP when the PCA is applied. This pattern of behavior is not followed by other ML regressors such as the SVR.

5. Discussion

Storage hydro-power generation mitigates the productivity uncertainty related to the intermittency of renewable resources and offers flexibility to the power system. On the contrary, in the medium and long-terms, uncertainty concerning meteorological and climatic processes affects hydro-power generation despite storage technology. For this reason, the individual operators base their planning on the comparison of current and future scenarios of power market prices and production capacity. As this is a common behavior for all hydro-power producers, the expected production capacity per market zone influences the related hydro-power supply. For instance, expectations of low hydro-power production capacity can induce operators to store resources in order to use them during more profitable periods, and this could cause a decrease in supply reflected in prices. From this point of view, prediction of hydro-power production capacity provides financial operators with a valid support for portfolio decision making. In addition, as the water resources of some reservoirs are used for more than one purpose, conflicts may arise in the strategic planning process to define priorities of use (e.g. on the balance among energy, environment and water management priorities). For these reasons, being able to predict hydro-power production capacity on a monthly basis can be a useful support to policy makers who may have the information they need to act promptly, both in terms of energy policies (i.e., to maintain the share of renewables and thus coordinate solar–wind–water strategies) and in terms of political strategies (i.e., strategic planning).

The main difficulty in constructing a model for the prediction of hydro-power production capacity is to include the complex phenomenon of snowpack melting, which represents the main driver of hydro-power production capacity, and has its highest intensity in the spring months, when most of the snowpack turns into an available water resource. In use case we have shown the existence of a high level of correlation among the input variables considered for the problem. This information redundancy suggests the possibility of reducing the number of input variables by means of feature reduction techniques, such as principal component analysis and variable grouping techniques. In fact, Figs. 7 and 9 show that the prediction error of the production capacity is dramatically reduced when, instead of considering all the variables, we apply a data reduction using the grouping proposed in Section 4.2.

Specifically, among all the tested methods, the multi-layer perceptron applied to the grouped variables has the best levels of MAE. In particular, it captures very well the trend of production capacity. This is corroborated by the fact that the value of the correlation between the actual production capacity and its prediction is very high (0.9735). This indicates that a feature reduction helps

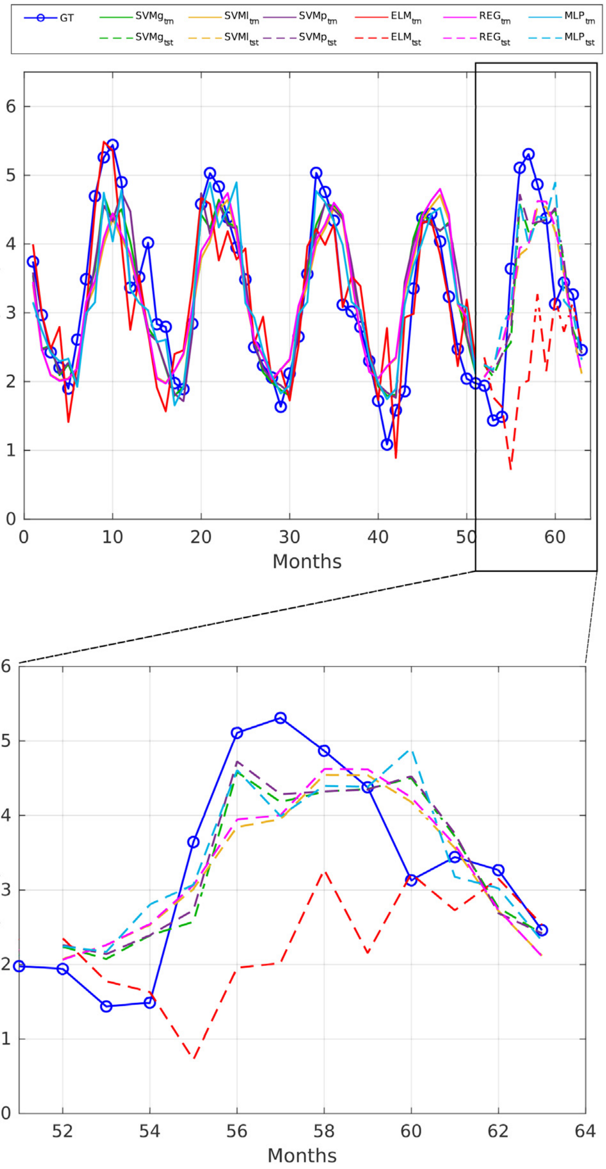


Fig. 10. Results of the considered ML regression techniques by fixing PCA components to 2 and using grouped variables.

this regressor improve its performance. In this connection, our results show that, in the case of grouped variables, by reducing the number of variables with the PCA technique, the quality of the MLP prediction gets worse, both in terms of MAE and of trend. This behavior is not observed in the performance of the SVR algorithm, which is improved in all cases where a feature reduction process is carried out.

Let us remark that grouping the variables related to each plant by sub-basins provides a physical foundation to the model. This is due to the input variables are calculated by taking into account only the zones in which the plants are allocated, and the grouping is done on the basis of a natural territorial contiguity. This ensures that policy makers or financial operators receive accurate information about the actual hydro-power production capacity for that market zone.

Moreover, MAE values obtained by the MLP are low in the months that are crucial for the snow-melting process, compared to other models applied. This is a very important result, since the prediction is very accurate at the time of year when the snowpack has highest intensity of melting and therefore has a crucial influence on the production capacity.

Besides, observing Fig. 9, the MLP regressor, applied to the grouped variables, makes a larger prediction error at the time of year when rainfall is most relevant for hydro-power production capacity, with respect to snowpack melting. Unlike other weather events, rainfall is considered in the prediction model by using two variables, i.e., the Alps-Italy and the rest of Northern Italy. This is because rainfall, differently from snowpack, is distributed in a very unpredictable way. As a consequence, the data recorded by each rain gauge have to be processed to obtain an estimate of the rainfall flowing into the streams associated with each reservoir. Thus, if the data are available, it would be better to use zonal rainfall estimated so as to capture only the rainfall flowing into the plants.

Finally, the best prediction model proposed in this paper (formed by a MLP regressor using grouped variables without PCA) is a very useful tool for policy makers and financial operators because it correctly indicates the trend of future hydro-power production capacity. Specifically, the results indicate whether it will be lower or higher than the current one, and also provide a good prediction of the intensity of this variation.

6. Conclusions

This paper proposes a methodology for hydro-power capacity prediction based on machine learning regression techniques. These techniques are used for modeling the relationship between the meteorological and climatic variables and the hydro-power production capacity. We assess the performance of artificial neural networks, such as multi-layer perceptron (MLP), extreme learning machines (ELM), and also statistical learning algorithms such as support vector regression (SVR) algorithms with different kernels, linear, polynomial or Gaussian. We have also proven the existence of a high level of correlation among the input variables considered for the problem. This information redundancy suggests the possibility of reducing the number of input variables by means of feature reduction techniques, such as principal component analysis (PCA) and variable grouping techniques.

We have tested the proposed prediction system in a real problem of monthly hydro-power capacity prediction in Northern Italy. Specifically, the MLP obtains the best result in the problem with a 0.2593 and 0.2128 TWh of root mean square error and mean absolute error, respectively, and a correlation of 0.9735 using the grouped variables which clearly outperforms all other evaluated methods. However, we also observe that the MLP does not improve its results when applying PCA. This behavior suggests that the MLP is able to process the information of the input variables better than the other regression algorithms, which improves their results when a PCA is applied (less information is available), but still have results far from those obtained by the MLP with the original input set.

Table A.5
Summary of the main acronyms used in the paper.

Acronym	Expansion
ML	Machine learning
MLP	Multi-layer perceptron
ELM	Extreme learning machine
SVR	Support vector regression algorithm
PCA	Principal component analysis
REG	Linear regression algorithm
CORR	Pearson correlation coefficient
MAE	Mean absolute error
RMSE	Root Mean Square Error

Table C.6
Results of the ML regression techniques using all the input variables considered. Error values are computed over the test set (the twelve months of the year 2018).

Method	CORR	RMSE	MAE
SVRg	0.6338	**1.1308**	**0.9177**
ELM	**0.6636 ± 0.1446**	1.2189 ± 0.3934	0.9511 ± 0.3345
REG	−0.3836	15.6946	11.1918
MLP	0.5774 ± 0.0540	1.8093 ± 0.4598	1.4246 ± 0.1227

The results obtained also indicate that using meteorological variables referring to each hydro-power plant seems to be better than considering all the meteorological data available for the entire market zone, for this specific problem. The proposed methodology allows considering the effect of complex phenomena, such as snowpack melting, in a model for the prediction of hydro-power production capacity. In addition, grouping the variables by sub-zones based on hydrographic sub-basins, ensures that the policy maker or the financial operator receives information on the actual hydro-power production capacity for that market zone.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

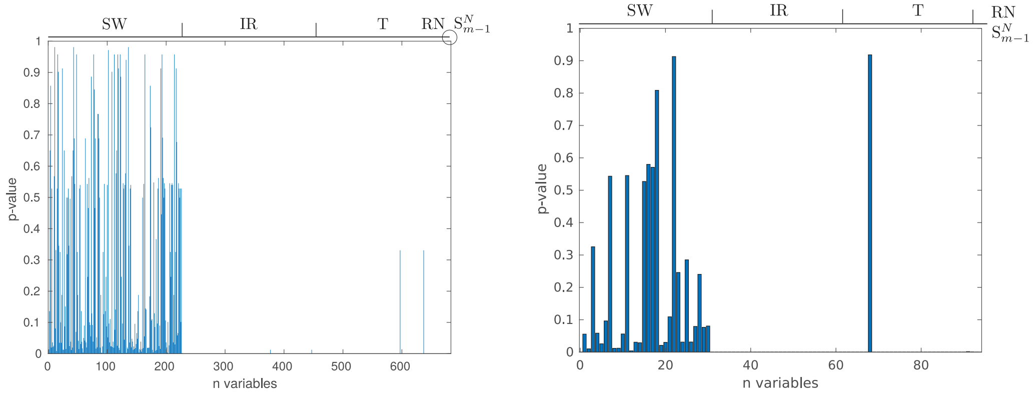
This research has been partially supported by Ministerio de Economía, Industria y Competitividad of Spain (grant ref. TIN2017-85887-C2-2-P) and by Comunidad de Madrid, Spain, PROMINT-CM project (grant ref: P2018/EMT-4366).

Appendix A. Acronyms

Table A.5 summarizes the acronyms used along the paper.

Appendix B. Metrics

We use the following metrics for giving quantitative results in the experimental part of the paper: the Root Mean Square



(a) All input variables vs output variable.

(b) Grouped variables vs output variable.

Fig. D.11. P-values of whether the correlation is significantly different from zero.

Error (RMSE), the Mean Absolute Error (MAE) and the Newman–Pearson coefficient, also known as Correlation Coefficient (CORR), which have the following formulae:

$$\begin{aligned}
\text{RMSE} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|y_i-\tilde{y}_i\|^2} \\
\text{MAE} &= \frac{1}{N}\sum_{i=1}^{N}|y_i-\tilde{y}_i| \qquad \text{(B.1)} \\
\text{CORR} &= \frac{\sum_{i=1}^{N}(y_i-E[y_i])(\tilde{y}_i-E[\tilde{y}])}{\sqrt{\sum_{i=1}^{N}(y_i-E[y_i])^2}\sqrt{\sum_{i=1}^{N}(\tilde{y}_i-E[\tilde{y}])^2}},
\end{aligned}$$

where y_i is ground truth, $\tilde{y}_i$ is the estimation and $E[\cdot]$ refers to the sample mean. RMSE and MAE tend to zero with good predictions and CORR approximates to 1 where ground truth and prediction are correlated. We refer the reader to the book [43] for a more detailed description of the usual ML evaluation metrics.

Appendix C. Tables

Table C.6 shows the results of the ML regression techniques using all the input variables (Section 4.5).

Appendix D. Statistical hypothesis testing for correlation values

We develop a statistical hypothesis testing to decide whether the correlation between the input variables and the output variable are significantly different from zero. Fig. D.11 shows the corresponding p-values of this testing.

Remember that p-values lower than 0.1 guarantee non-null correlations between the input and the output considered which a significant probability. As it can be seen, the p-values associated with the rainfall, solar radiation and temperature are by far lower than 0.1. Most of the snowpack depth variables are also lower than this reference value. However, there are some of them which are above it. The reason is that these variables correspond to areas where the presence of snow is not significant for the hydro-power production capacity prediction.

References

[1] M.H. Albadi, E.F. El-Saadany, Overview of wind power intermittency impacts on power systems, Electric Power Syst. Res. 80 (2010) 627–632.

[2] R. Komiyama, Y. Fujii, Assessment of massive integration of photovoltaic system considering rechargeable battery in Japan with high time-resolution optimal power generation mix model, Energy Policy 66 (2014) 73–89.

[3] B. Hamududu, A. Killingtveit, Assessing climate change impacts on global hydropower, Energies 5 (2012) 305–322.

[4] IEA, International Energy Agency, https://www.iea.org/.

[5] P. Denholm, M. Hand, Grid flexibility and storage required to achieve very high penetration of variable renewable electricity, Energy Policy 39 (3) (2011) 1817–1830.

[6] M. K.Chang. J. D. Eichman, F. Mueller, S. Samuelsen, Buffering intermittent renewable power with hydroelectric generation: A case study in california, Appl. Energy 112 (2013) 1–11.

[7] M.S. Javed, T. Ma, J. Jurasz, M.Y. Amin, Solar and wind power generation systems with pumped hydro storage: Review and future perspectives, Renew. Energy 148 (2020) 176–192.

[8] K.H. Chang, A decision support system for planning and coordination of hybrid renewable energy systems, Decis. Support Syst. 64 (2014) 4–13.

[9] P.D. Lund, J. Lindgren, J. Mikkola, J. Salpakari, Review of energy system flexibility measures to enable high levels of variable renewable electricity, Renew. Sustain. Energy Rev. 45 (2015) 785–807.

[10] L. Hirth, The benefits of flexibility: The value of wind energy with hydropower, Appl. Energy 181 (2016) 210–223.

[11] J. Després, S. Mima, A. Kitous, P. Criqui, N. Hadjsaid, I. Noirot, Storage as a flexibility option in power systems with high shares of variable renewable energy sources: a POLES-based analysis, Energy Econ. 64 (2017) 638–650, The Journal of Energy Markets' Special Issue, 14.1 2021.

[12] M.W. Tian, S.R. Yan, X.X. Tian, S. Nojavan, K. Jermsittiparsert, Risk and profit-based bidding and offering strategies for pumped hydro storage in the energy market, J. Cleaner Prod. 256 (2020) 120715.

[13] A.R. Kian, J.B. Cruz, Bidding strategies in dynamic electricity markets, Decis. Support Syst. 40 (3–4) (2005) 543–551.

[14] C. Condemi, L. Mastroeni, P. Vellucci, Selection of predictor variables to aggregate generation model, J. Energy Mark. 14.1 (2021).

[15] Ellen Krohn Aasgård, et al., Hydropower bidding in a multi-market setting, Energy Syst. 10.3 (2019) 543–565.

[16] R. Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future, Int. J. Forecast. 30.4 (2014) 1030–1081.

[17] Reston Filho, et al., Energy price prediction multi-step ahead using hybrid model in the Brazilian market, Electr. Power Syst. Res. 117 (2014) 115–122.

[18] Claudio Monteiro, L. Alfredo Fernandez-Jimenez, Ignacio J. Ramirez-Rosado, Explanatory information analysis for day-ahead price forecasting in the iberian electricity market, Energies 8.9 (2015) 10464–10486.

[19] B. Plucinski, Y. Sun, S.Y. Wang, R.R. Gillies, J. Eklund, C.C. Wang, Feasibility of multi-year forecast for the colorado river water supply: Time series modeling, Water 11 (2019) 2433.

[20] H. Pan X. Lv, Reconstruction of spatially continuous water levels in the columbia river estuary: The method of empirical orthogonal function revisited, Estuar. Coast. Shelf Sci. 222 (2019) 81–90.

[21] X. Zhang, P. Liu, Y. Zhao, C. Deng, Z. Li, M. Xiong, Error correction-based forecasting of reservoir water levels: Improving accuracy over multiple lead times, Environ. Model. Softw. 104 (2018) 27–39.

[22] P. Goovaerts, Geostatistical prediction of water lead levels in flint, michigan: A multivariate approach, Sci. Total Environ. 647 (2019) 1294–1304.

[23] R.R. Karri, X. Wang, H. Gerritsen, Ensemble based prediction of water levels and residual currents in Singapore regional waters for operational forecasting, Environ. Model. Softw. 54 (2014) 24–38.

[24] B. Bazartseren, G. Hildebrandt, Short-term water level prediction using neural networks and neuro-fuzzy approach, Neurocomputing 55 (2003) 439–450.

[25] F.J. Chang, Y.T. Chang, Y.T. Adaptive neuro-fuzzy inference system for prediction of water level in reservoir, Advances in Water Research 29 (2006) 1–10.

[26] A.P. Wang, H.Y. Liao, T. Chang, Adaptive Neuro-fuzzy Inference System on Downstream Water Level Forecasting. in: Proceedings of the 2008 IEEE Fifth International Conference on Fuzzy Systems and Knowledge Discovery, Shandong, China, 18–20 October, 3 (2008) 503-507.

[27] D. Zhang, J. Lin, Q. Peng, D. Wang, T. Yang, S. Sorooshian, X. Liu, J. Zhuang, Modeling and simulating of reservoir operation using the artificial neural network, support vector regression, deep learning algorithm, J. Hydrol. 565 (2018) 720–736.

[28] S. Yang, D. Yang, J. Chen, B. Zhao, Real-time reservoir operation using recurrent neural networks and inflow forecast from a distributed hydrological model, J. Hydrol. 579 (2019) 124229.

[29] C. Castillo-Botón, D. Casillas-Pérez, C. Casanova-Mateo, L.M. Moreno-Saavedra, B. Morales-Díaz, J. Sanz-Justo, P.A. Gutiérrez, S. Salcedo-Sanz, Analysis and prediction of dammed water level in a hydropower reservoir using machine learning and persistence-based techniques, Water 12 (2020) 1528.

[30] N. Chen, C. Xiong, W. Du, C. Wang, C. Lin, Z. Chen, An improved genetic algorithm coupling a back-propagation neural network model (IGA-BPNN) for water-level predictions, Water 11 (2019) 1795.

[31] S. Samadianfard, S. Jarhan, E. Salwana, A. Mosavi, S. Shamshirb, S. Akib, Support vector regression integrated with fruit fly optimization algorithm for river flow forecasting in lake urmia basin, Water 11 (2019) 1934.

[32] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, 1998.

[33] C.M. Bishop, Neural Networks for Pattern Recognition, Oxford University Press, 1995.

[34] G.B. Huang, Q.Y. Zhu, Extreme learning machine: theory and applications, Neurocomputing 70 (2006) 489–501.

[35] A.J. Smola, B. Schölkopf, A tutorial on support vector regression, Stat. Comput. 14 (2004) 199–222.

[36] S. Salcedo-Sanz, J.L. Rojo, M. Martínez-Ramón, G. Camps-Valls, Support vector machines in engineering: an overview, WIREs Data-Mining Knowl. Discov. 4 (3) (2014) 234–267.

[37] M.T. Hagan, M.B. Menhaj, Training feed forward network with the marquardt algorithm, IEEE Trans. Neural Netw. 5 (6) (1994).

[38] G.B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern. B 42 (2) (2012) 513–529.

[39] D.A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press, 2009.

[40] G.B. Huang, ELM matlab code, http://www.ntu.edu.sg/home/egbhuang/elm_codes.html.

[41] S.J. Axler, Linear Algebra Done Right, Springer, New York, 1997, (2).

[42] ISTAT, Istituto Nazionale di Statistica, https://www.istat.it/.

[43] Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[44] H.M. Zawba, E. Emary, C. Grosan, V. Snasel, Large-dimensionality small-instance set feature selection: A hybrid bio-inspired heuristic approach, Swarm Evol. Comput. 42 (2018) 29–42.