



# Piezometric level prediction based on data mining techniques

Joaquim Tinoco<sup>1</sup> · Mathilde de Granrut<sup>2,3</sup> · Daniel Dias<sup>3</sup> · Tiago Miranda<sup>1</sup> · Alexandre-Gilles Simon<sup>2</sup>

Received: 9 March 2018 / Accepted: 18 July 2019 / Published online: 1 August 2019  
© Springer-Verlag London Ltd., part of Springer Nature 2019

## Abstract

The safety assessment of dams is a complex task that is made possible thanks to a constant monitoring of pertinent parameters. Once collected, the data are processed by statistical analysis models in order to describe the behaviour of the structure. The aim of those models is to detect early signs of abnormal behaviour so as to take corrective actions when required. Because of the uniqueness of each structure, the behavioural models need to adapt to each of these structures, and thus flexibility is required. Simultaneously, generalization capacities are sought, so a trade-off has to be found. This flexibility is even more important when the analysed phenomenon is characterized by nonlinear features. This is notably the case of the piezometric levels (PL) monitored at the rock–concrete interface of arch dams, when this interface opens. In that case, the linear models that are classically used by engineers show poor performances. Consequently, interest naturally grows for the advanced learning algorithms known as machine learning techniques. In this work, the aim was to compare the predictive performances and generalization capacities of six different data mining algorithms that are likely to be used for monitoring purposes in the particular case of the piezometry at the interface of arch dams: artificial neural networks (ANN), support vector machines (SVM), decision tree,  $k$ -nearest neighbour, random forest and multiple regression. All six are used to analyse the same time series. The interpretation of those PL permits to understand the phenomenon of the aperture of the interface, which is highly nonlinear, and of great concern in dam safety. The achieved results show that SVM and ANN stand out as the most efficient algorithms, when it comes to analysing nonlinear monitored phenomenon. Through a global sensitivity analysis, the influence of the models' attributes is measured, showing a high impact of  $Z$  (relative trough) in PL prediction.

**Keywords** Dam monitoring · Concrete dams · Piezometric levels · Data mining techniques

## 1 Introduction

### 1.1 Context

Assessing the safety of dams is a priority for the owners of those large civil engineering structures. They are required to have a clear vision of the state of health of their dams,

and to be able to potentially detect any abnormal evolution. In case such an event occurs, identifying the causes and taking the necessary steps to bring the structure back to a safe state is made possible thanks to a good knowledge and understanding of the behaviour of the structure.

Considering the specificity of each structure (the geometrical parameters, the hydro-geological context, the shape of the valley in which it is situated, specific operational conditions, the historical background in terms of environmental conditions, climatic events...), being able to assess at any moment the state of a given dam is a complex task, that relies on a systematic surveillance of the structure. This surveillance is based on the one hand on visual inspection, and on the other hand on monitoring. While visual inspections are rather qualitative techniques, the monitoring of dams is based on the continuous gathering of pertinent measurements that are processed by behavioural analysis models. The type of data that are collected is

✉ Joaquim Tinoco  
jtinoco@civil.uminho.pt

<sup>1</sup> ISISE - Institute for Sustainability and Innovation in Structural Engineering, School of Engineering, University of Minho, Guimarães, Portugal

<sup>2</sup> Électricité de France - Division Technique Générale (EDF DTG), Grenoble, France

<sup>3</sup> 3SR Laboratory, CNRS, Grenoble INP, Grenoble Alpes University, Grenoble, France

varied, including mechanical (displacements) and hydraulic quantities (piezometric levels, leakage flows). Those quantities constitute representative factors that traduce the global behaviour of the dam. Consequently, they are analysed to describe as finely as possible this behaviour. In engineering practices, the behaviour of the dam is assumed to be simultaneously influenced by three external loads, namely the hydrostatic load, the thermal load, and the temporal load. Those loads are thus also measured. Eventually, behavioural models are built, based on the measurements of both loads and their effects [10, 22, 24, 26].

The first aim of those models is to provide a prediction of the structure behaviour under normal operating conditions, which is compared to the actual measurements and makes it possible to check the appropriateness between the expected and actual evolutions. Second, in a long-term perspective, sensitivity analyses are carried out to identify the contribution of each load to the monitored parameters, which permit to assess the overall functioning of the structure.

Today, as most dams have been monitored since their first filling, and with the generalization of telemetry, a great amount of data is already available, which makes it possible to use statistical models. Among the community of dam owners, the classically used models belong to the category of the multilinear regression, and the reference model is the HST (Hydrostatic, Season, Time) model [41]. Initially designed to describe mechanical phenomena, this type of model assumes that the explanatory factors have independent and thus additive effects on the modelled quantity. Thus, its application to the analysis of hydraulic phenomenon is not always pertinent, for nonlinearity comes into play, and the additivity assumption is invalidated. In order to deal with such nonlinear phenomena, more advanced models issued from the data mining techniques turn out to be particularly interesting, by providing valuable processing of the database.

## 1.2 The aperture of the rock–concrete interface: a nonlinear phenomenon

The state of compression of the rock-mass foundation situated right under the contact between rock and concrete is in constant evolution, due to the variations of the abutment forces that the foundations support. The appearance of tensile stress is regularly observed at the heel of the dam, which causes the permeability of the rock-mass to increase, and the hydrostatic load is thus transferred to the foundation. The tensile stress can also induce an unsticking of the rock–concrete contact, and/or a cracking of the upstream face concrete [21].

This phenomenon is referred to as the opening of the rock–concrete interface, considering the interface as the

1-m-wide zone on both sides of the contact. The aperture of the contact (Fig. 1) induces the development of uplift pressure, that is to say the rise of the piezometry in that zone. Subsequently, it is possible to assess the state of aperture by interpreting the piezometric levels measured at the interface.

This aperture is irreversible, for rock cannot cicatrize by itself. Its temporal evolution is multiscale: indeed, the size of the aperture varies at the infra-annual scale, evolving with the mechanical stress that the dam is subject to, but its magnitude can also evolve at the scale of several years, with the opening, and thus the full charge propagating further towards the toe of the dam, or on the contrary, declining subsequently to specific operational conditions, clogging, etc.

Unlike most mechanical phenomena, the aperture of the interface follows nonlinear evolution rules. Indeed, because of the thermal sensitivity of concrete, the influence of a given filling (and thus of a given hydrostatic load) differs according to the thermal state of the structure: low temperatures cause concrete to contract, inducing a global downstream movement of the arch, which exacerbates the tensile stress and finally increases the aperture. Conversely, high temperatures tend to “close” the aperture. Thus, this phenomenon has to be studied taking into account complex interactions between the different influence quantities.

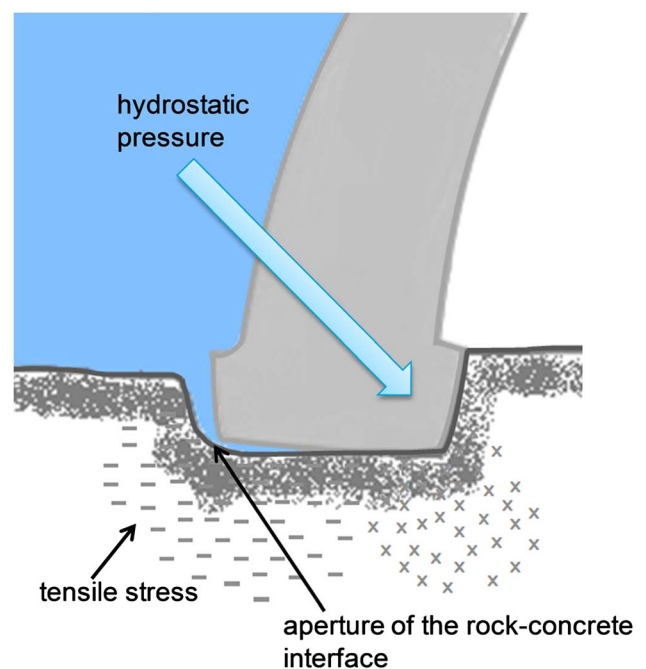


Fig. 1 Rock–concrete aperture

### 1.3 Motivation and objectives

Because of the nonlinear features that characterize the piezometry at the rock–concrete interface, it is not possible to obtain a satisfying modelling by using mere additive models. Consequently, a more advanced type of model was sought among the different data mining (DM) techniques, the use of which is growing in Civil Engineering, and more particularly in dam monitoring. With the by now diversified range of existing DM techniques, the identification of the most pertinent algorithm is not obvious. Although there is an extensive literature presenting studies that were conducted in the field of dam monitoring, when focusing on arch dams, the majority of them deal with mechanical phenomena (vertical, radial and, tangential displacements). To pick one example out, the subject of theme C at the 6th ICOLD Benchmark Workshop on Numerical Analysis of the Dams [18] was the analysis of the crest displacement of the Schlegeis arch dam (Austria). Various deterministic and statistical models and combinations of them were used, namely the multilinear regression (MLR), autoregressive moving average (ARMA), finite elements (FE), nonlinear autoregressive with exogenous input (NARX), artificial neural networks (ANN), trial load method (TLM) and nonparametric polynomials (NP). This case study was then resumed by Balcilar and Demirkaya [2] who applied the MLR and ANN models. They came to the same conclusion, which is that linear models are perfectly suitable to interpret displacement measurements. Those two references are only a sample of the numerous applications of MLR models and more sophisticated algorithms to the analysis of displacements. Many other studies are available, among which Mata [23], Ranković et al. [28] and Rankovic et al. [27, 29]. As recalled by Balcilar and Demirkaya [2], mechanical phenomena are in most cases purely additive. Therefore, DM techniques do not provide a tremendous improvement compared to linear models and even sometimes show worst performances [2].

Conversely, while modelling hydraulic phenomena (piezometry, leakage) with such advanced methods would benefit from the high adaptability of those techniques, and their capacity to adapt to complex interactions between the inputs, leakage and piezometry measurements are actually seldom addressed. A recent exhaustive review was done by Salazar et al. [32, 34] which draws up an inventory of the statistical and machine learning techniques that have been applied to dam behaviour modelling, and discusses the methodological choices that were made by the authors. No less than 59 studies were noted and analysed, each study corresponding to one type of output (radial, tangential or vertical displacement, leakage, piezometry, etc.) collected on one or more dams, and analysed with one or more

techniques. The techniques that were addressed are the following: MLR, impulse response function (IRF),  $k$ -nearest neighbours (KNN), ANN, wavelet neural networks (WNN), nonlinear autoregressive exogenous neural network (NARXNN), autoassociative neural network (AANN), NARX, Hybrid (HYB), adaptive neurofuzzy system (ANFIS), principal component analysis (PCA), moving PCA (MPCA), support vector machines (SVM), error correction model (ECM), robust regression (RR), multivariate adaptive regression splines (MARS), random forest (RF) and boosted regression trees (BRT). Among the reviewed studies, 38 deal with radial displacements, 31 of which in arch dams. Hydraulic indicators (piezometry, leakage), however, are much less taken into consideration, especially in arch dams, although considerable safety issues are at stake, as shown by the matter of the foundation uplift pressures. Piezometry in arch dams is addressed in only one out of the 59 studies [16], using a MLR model.

Apart from this review paper, the state of the art confirms the scarcity of references relative to arch dams piezometry and leakage measurements. Simon et al. [36] analysed piezometric measurements with ANN. They put forward the capacity of ANN to detect the nonlinear links between the input variables, which permits much finer and more accurate predictions compared to MLR analysis. This paper is the only one that could be found addressing the triptych “piezometry—arch dam—DM techniques”. Rankovic et al. [27, 29] propose an application of ANN to the analysis of piezometric levels but in an embankment dam. This type of dam behaves very differently compared to concrete dams, and the nonlinear features that are the subject of the present study are not at stake. In addition, the variables that are used as inputs do not include time. This implies that no irreversible evolution is taken into account, which is not acceptable in dam monitoring.

Leakage measurements share more common features with piezometric levels, since leakages that are collected in the rock–concrete interface are also influenced by the aperture of the interface. Consequently, one can expect the same nonlinear features to be found when analysing such types of measurements. When enlarging the bibliographical research to leakages, Simon et al. [36] also propose an application of ANN, which help identify the link between the inputs and the output, and lead to the improvement of a state-of-the-art linear regression model. In a comparative work, Salazar et al. [32, 34] benchmark different machine learning techniques applied to dam behaviour analysis. They apply models based on MLR, ANN, SVM, MARS, RF and BRT techniques to the La Baells arch dam leakage and displacement measurements. Those leakage measurements include measurements that were made in the central block of the dam, and that are thus potentially influenced by an aperture of the interface. The models are altogether

adjusted to 14 target variables, using 25 predictors. The comparison of the different results obtained is made on the basis of the mean absolute error (MAE), and the average relative variance (ARV), which is the ratio between the mean squared error and the variance. As far as leakages are concerned, what comes out of this study is that the BRT are the most performing models. However, as underlined by the author, all techniques are not equivalent in terms of tuning effort, depending on the number of parameters by which they are defined, their flexibility, and their sensitivity to noise. Consequently, in this study that seeks to be as unbiased as possible, the modelling choices that are made necessarily induce that each technique is not driven to its maximum performance. This shows particularly well when looking at the results corresponding to the leakages collected on the central block. For those two targets, while the MLR models scores a MAE of 2.6 mm, ANN and SVM, respectively, score 3.04 mm and 5.38 mm. This shows that SVM and ANN could be improved by a more detailed tuning and as suggested by the author, a more careful variable selection. Consequently, BRT might not actually be the most performing algorithm to analyse leakage time series, so further tests have to be conducted.

Finally, Santillán et al. [35] propose an application of ANN to analyse seepages on an arch dam. A sensitivity analysis is proposed to visualize how each input impacts the seepage. However, the author develops an algorithm aimed at selecting automatically the input variables among the following possibilities: the air temperature, the water level in the reservoir, and several moving averages of both variables. In the end, the retained inputs are the water level and two moving averages of this water level, which implies that no thermal influencing quantity is considered, nor any irreversible evolution. This actually contradicts what is actually observed on all French arch dams.

From this state of the art, it seems that as far as PL predictions are concerned, ANNs are the most popular algorithm. They are, however, often regarded as a “black box”, and no agreement is reached regarding the pre-processing of the data, the way to build the learning and test sets, how to choose the architecture of the network, the stopping criteria, the number of iterations, or how to check the generalization capacities. Some authors even use this algorithm for temporal extrapolation [27, 29], while others strongly advice against it [6, 33].

Consequently, the following work aims at filling the above-mentioned gaps by performing a comparative analysis of six of the most popular DM techniques in civil engineering applied to the monitored arch dam piezometric levels. Those are artificial neural networks (ANN), support vector machines (SVM), decision tree (DT),  $k$ -nearest neighbour (kNN), random forest (RF) and multiple regression (MR). The aim is to get a better understanding

of how the issue of the aperture of the rock–concrete interface can be analysed, which tool is the most suitable for that purpose. To that end, the prediction performances of the six methods are assessed and compared. Eventually, the interpretability of those complex algorithms is declined thanks to a global sensitivity analysis procedure.

## 2 Case study (dam and data)

### 2.1 The dam

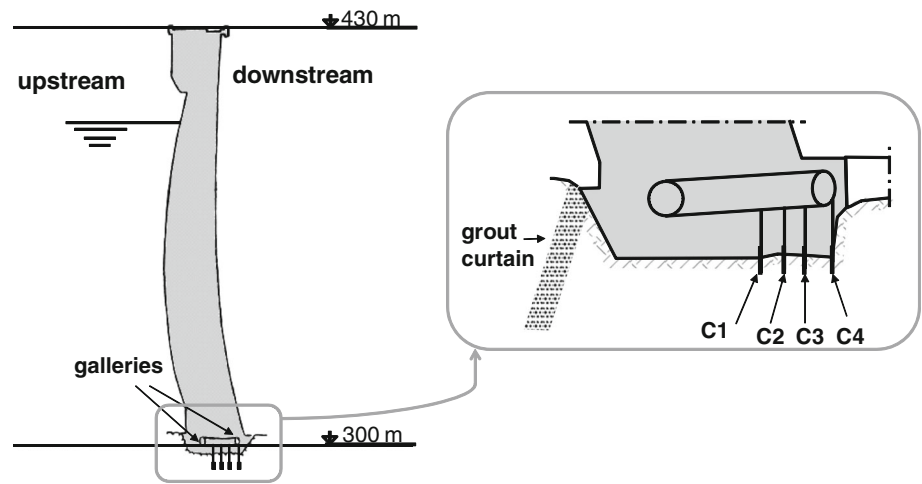
The data that were proposed for that study comes from a French double-curvature arch dam, which is 130 m high from the foundation to the crest, with a 425-m-long crest. Its thickness varies from 25 to 6 m, and it is thus considered a thin arch. The ratio between the width of the valley ( $L$ ) and the height of the dam ( $H$ ) is higher than 3 ( $L/H = 3.3$ ), which indicates that the valley is relatively large. Those characteristics are elements that are known to favour the appearance of an aperture at the interface. Coupled to the phenomena of shrinkage and creep of concrete and the creep of the foundation, the arch shifted to the downstream direction right from the first filling of the reservoir, and the rock–concrete interface opened (5–7 mm). Subsequently, uplift pressures propagated under the heel of the dam, and further toward downstream. In order to follow the evolution of this phenomenon, the network of piezometers has been gradually complemented. It now comprises 52 piezometers, distributed under the dam and under the downstream plunge basin. Among those sensors, four are particularly interesting (named C1, C2, C3 and C4), since they are situated (Fig. 2) in the central cantilever of the dam, at the rock–concrete interface (298 m). Therefore, they are directly influenced by the aperture of this interface. The data provided for this study corresponds to the C4 sensor, situated close to the downstream end of the toe.

C4 is the piezometer that is situated at the smallest distance from the downstream end. Therefore, assuming that the aperture extents until the toe when the dam is submitted to extremely high stresses, C4 is situated in a zone where the interface is sometimes closed (for low fillings and/or high temperatures) and sometimes open (for high water levels and/or low temperatures). Thus, C4 is the most sensitive to the load variations and can be used to get an idea of the evolution of the aperture.

### 2.2 The data

The time series that are provided stretch from September 2011 to June 2016 and comprise 623 observations for each measured quantity. The measured quantities are the

**Fig. 2** Piezometers location (central cantilever)



following, with  $i \in \{1623\}$  (all measurements are synchronized):

- the piezometric levels  $PL_i$  expressed in meters of a water column
- the water level in the reservoir  $h_i$  expressed in meters
- time  $t_i$  expressed in number of days elapsed since 1 January 2011 (the chosen origin)

The following variables are subsequently defined from those measurements:

- the season  $S_i$  which is an angle equal to  $0^\circ$  on the 1st of January and  $360^\circ$  on the 31st of December, defined by  $S_i = 2\pi \cdot \left(\frac{t_i}{365.25} - \text{floor}\left(\frac{t_i}{365.25}\right)\right)$ , where the function floor() computes the integer part of its argument.
- the relative trough  $Z_i$  which is a scaling of the water level  $h_i$  defined by  $Z_i = \frac{h_{\text{norm}} - h_i}{h_{\text{norm}} - h_{\text{emp}}}$ , where  $h_{\text{norm}}$  is the normal operating water level and  $h_{\text{emp}}$  the water level when empty

In order to have a temporal distribution that would be as balanced as possible, a time sampling was processed so as to keep maximum one measurement per day. Indeed, in standard operating conditions, the sensors are automatically polled once a week during the hot season and/or for low water level, and once or twice a day during cold season and for high water levels. However, during some singular operational events, the data acquisition frequency increases in order to conduct the operation as safely as possible and follow its evolution closely. For the considered sensor, the interval between two measurements is often inferior to 2 days, and up to ten measurements per day are regularly observed (twice a year, the drainage system is locally closed and opened for efficiency reasons). Consequently, those very close measurements are correlated between each other, and the convergence of the models might be overly influenced by those identical observations. What is more,

the fact that those dense observations correspond to singular operational conditions might result in a deterioration of the representativeness (keeping in mind that the models aim at describing the behaviour of the dam under normal operating conditions). Consequently, it was decided to keep only one measurement per day which resulted in 623 remaining observations (starting from an initial 1041-large dataset). No more advanced sampling was carried out, in order to keep a sufficient amount of observations.

Table 1 summarizes the main statistics of the variables used as model attributes as well as the PL. Figure 3 depicts the histograms of  $Z$  and  $h$ ,  $S$ , as well as PL.

### 3 Modelling

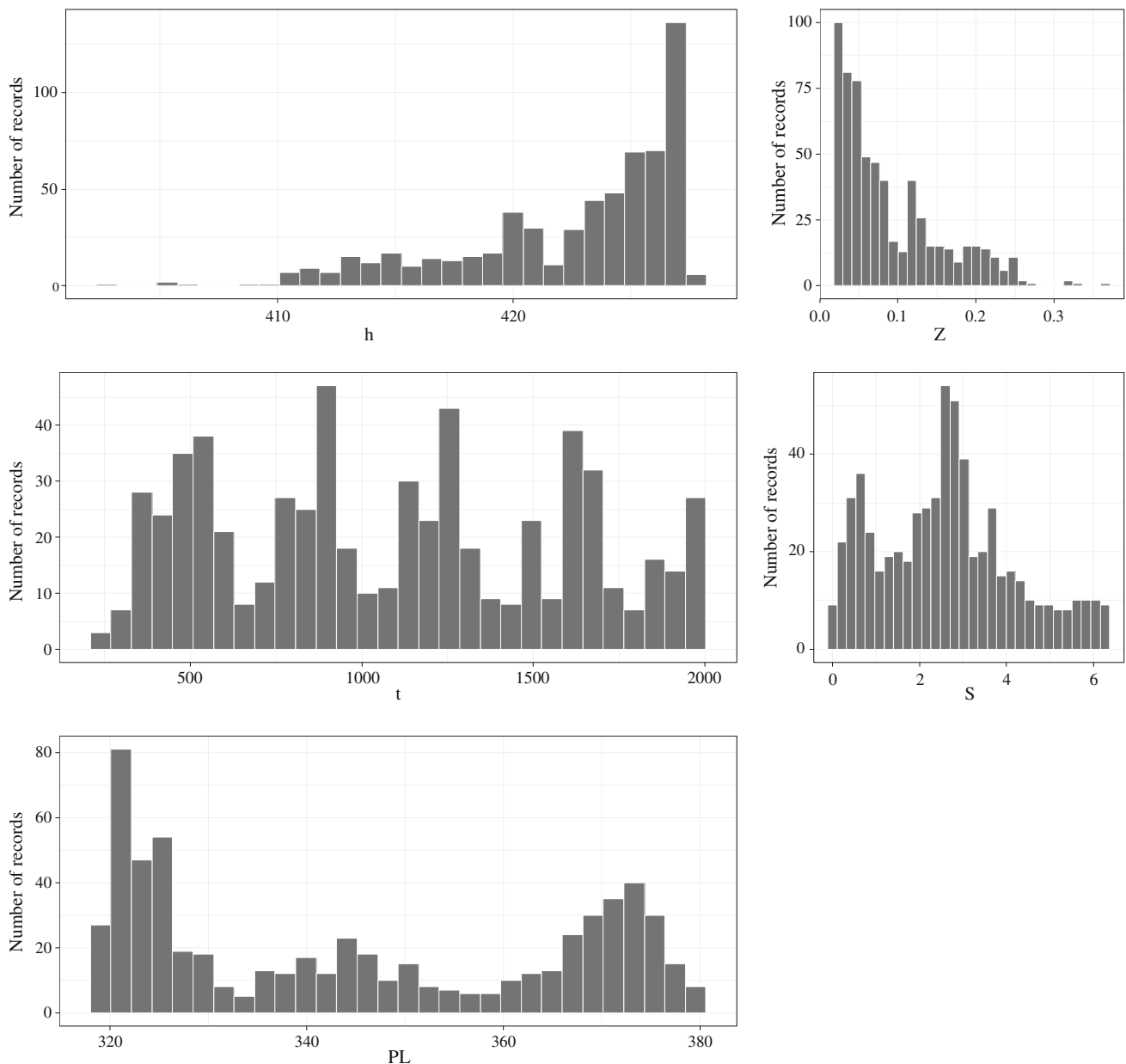
#### 3.1 Choice of the predictors

The modelling techniques that are compared for regression purposes are based on complex mathematical algorithms and do not take into account any physical law. Thus the choice of the input variables is a way to involve some physical understanding of the phenomenon at stake. In the present case, the algorithms are the tools that are used to model the link between the PL and the three main external loads. Each of these loads has an effect on the PL, and it is

**Table 1** Main statistics of the models attributes and output (PL)

Variable	Minimum	Maximum	Mean	SD
$Z$ (-)	0.0178	0.3599	0.0891	0.0653
$h$ (m)	402.69	427.7	422.48	4.78
$S$ (-)	0.0055	6.2698	2.6114	1.5584
PL (m)	318.8	397.3	345.3	21.1





**Fig. 3** Histograms of all models attributes and output (PL)

those effects that the models are expected to build adequately from the appropriate predictors, detecting nonlinear interactions between the inputs. The first load is the hydrostatic load, and the corresponding predictor is the relative through  $Z$ . The second load is the thermal load, which includes the different annual thermal waves that induce cyclical temperature variation in the concrete of the structure. This global concrete temperature variation is the sum of different thermal variations (air temperature as the most influent but also water temperature, solar radiations, presence or absence of wind, temperature of the foundations etc., with potentially different phases), that can thus be modelled thanks to periodic functions of the season,

with a 1-year-long period. The corresponding predictors are thus  $\cos(S)$  and  $\sin(S)$ . Transforming the season beforehand permits to define predictors that are as close as possible to the thermal load, as confirmed by the experience in monitoring air and concrete temperature. In addition, since the correlation between  $\cos(S)$  and  $\sin(S)$  is zero, using those variables as inputs does not induce the risk of multicollinearity. The last influence quantity is time, which is introduced into the regression through the variable  $t$ .

Engineering practices have established by custom the use of the historical model named Hydrostatic Season Time (HST) [41]. It is a multilinear regression model that uses additional predictors compared to those four basic ones.

Indeed, in order to take into account the influence of the thermal loads, not only are  $\cos(S)$  and  $\sin(S)$  introduced as inputs but also  $\cos(2S)$  and  $\sin(2S)$ . As to the effect of the hydrostatic load, a fourth-order polynomial in  $Z$  is used. Those choices were determined empirically. Considering  $X_i$  the measurement of a given phenomenon, the HST model is thus expressed as follows:

$$X_i = b_0 + f_1(t) + f_2(Z) + f_3(S) + \varepsilon_i, \quad i \in \{1, N\}$$

With  $f_1(t) = b_1 t$  the irreversible function,  $f_2(Z) = b_2 Z + b_3 Z^2 + b_4 Z^3 + b_5 Z^4$  the hydrostatic function,  $f_3(S) = b_6 \cos(S) + b_7 \sin(S) + b_8 \cos(2S) + b_9 \sin(2S)$  the seasonal function,  $\varepsilon_i$  the modelling residue (representing the measurement errors and model imperfections),  $b_0$  the constant of linear regression

For the following study, however, the objective is to compare the performances of the different algorithms by confronting them to the same problem, and thus the same inputs.

What is ideally expected from a model is to offer the necessary flexibility to be able to provide good predictions without having to impose physical laws a priori. Indeed, when modelling a complex phenomenon, the user is likely to identify a new potentially influencing variable, which he might want to add as a predictor, but without necessarily having a precise idea of how it affects the modelled quantity. Thus, the model is expected to build the interactions between this predictor and the modelled quantity automatically. That is the reason why it was chosen in this study to keep the inputs as basic as possible:  $Z$ ,  $t$ ,  $\cos(S)$  and  $\sin(S)$ , in order to identify the algorithm that has the highest flexibility and adaptability.

### 3.2 Used DM techniques

In the context of the monitoring of concrete arch dams, it is necessary to dispose of models that are able to describe as precisely as possible the phenomenon of cracking at the concrete-rock interface, leading to the development of uplift pressures, which are likely to threaten the stability of the structure.

Considering the low performance of the linear models that are classically used, this work intends to explore the capabilities of advanced statistical analysis, also known as data mining (DM) techniques. For that, six different DM algorithms were applied to analyse piezometric data monitored on a French large arch dam: artificial neural networks (ANN), support vector machines (SVM), decision tree (DT),  $k$ -nearest neighbour (kNN), random forest (RF) and multiple regression (MR). These advanced tools have been applied in different fields [4, 25], namely in civil engineering field [19, 39, 42], and take advantage of a consolidated experience.

ANNs is a computational model inspired by the structure and functions of biological neural networks [20]. It can be used for different purposes among which classification, pattern recognition, or, as in the case of that study, regression. Due to its high flexibility, it is capable of automatically detecting and modelling complex nonlinear relationships between its inputs and outputs.

A neural network is composed of different processing units called neurons. In the case of a multilayer perceptron, as used in this study, the neurons are organized in successive layers. In the first layer, known as the input layer, each neuron corresponds to one predictor. It receives the information and transfers it to each neuron in the following layer, called the hidden layer. In turn, those neurons process the data by computing the weighted summation of its inputs, and applying a function known as the transfer or activation function to this sum. The result is transferred to each neuron in the following layer. The last layer is the output layer, which produces the prediction. This propagation of the information through the network is the feedforward propagation. The parameters that need to be optimized on the learning set are the weights of the neurons. A retropropagation algorithm is used to carry out this optimization iteratively [30]. The more neurons a network contains, the more it will be able to fit the data. However, overfitting must be avoided in order to keep a satisfying generalization capacity. Consequently, a trade-off between fitting accuracy and generalization capability has to be found.

For that study, the R nnet package [38] was used. The network was designed with one hidden layer containing  $H$  neurons. In order to determine the adequate value for  $H$  (trade-off between fitting accuracy and generalization capability), a grid search of five values  $\{0; 2; 4; 6; 8\}$  was used during the learning phase. This grid search only considers training dataset (as defined by the threefold cross-validation methodology), dividing it randomly into fitting (70%) and validation data (30%), where the validation error was used to select the best  $H$ . After selecting the best  $H$  value, the ANN is retrained with the whole training dataset. The activation function of the hidden nodes was set to the popular logistic function  $1/(1 + e^{-x})$ . ANN optimization was done via the BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm [40]. The BFGS method is a quasi-Newton method (also known as a variable metric algorithm) that was published simultaneously in 1970 by Broyden, Fletcher, Goldfarb, and Shanno. This uses function values and gradients to build up a picture of the surface to be optimized [14]. As described in Cortez [14], the training (BFGS algorithm) is stopped when the error slope approaches zero or after a maximum of 100 epochs.

SVMs were initially proposed for classification tasks [12]. Then it became possible to apply SVM to regression tasks after the introduction of the  $\epsilon$ -insensitive loss function [37]. SVM has theoretical advantages over ANN, such as the absence of local minima in the learning phase. The main purpose of the SVM is to transform the input data into a higher dimensional space referred to as the feature space, using nonlinear mapping (see Fig. 4). The SVM then finds the best linear separating hyperplane ( $y_i = \omega_0 + \sum_{i=1}^m \omega_i \phi_i(x)$ ), related to a set of support vector points, in the feature space. The support vector points are the transformed observations that are the nearest to the hyperplane. The transformation  $\phi$  depends on a kernel function. In this work, the popular Gaussian kernel ( $k(x, x') = \exp(-\gamma \cdot \|x - x'\|^2), \gamma > 0$ ) was adopted, where  $x, x' \in \mathbb{R}^N$ . In this context, its performance is affected by three parameters:  $\gamma$ , the parameter of the kernel;  $C$ , a penalty parameter; and  $\epsilon$  (only for regression), the width of a  $\epsilon$ -insensitive zone [31], which defines a tolerance on the error. The heuristics proposed by Cherkassky and Ma [11] were used to define the last two parameter values,  $C = 3$  (for a standardized output) and  $\epsilon = \hat{\sigma} / \sqrt{N}$ , where  $\hat{\sigma} = 1.5 / N \cdot \sum_{i=1}^N (y_i - \hat{y}_i)^2$ ,  $y_i$  is the measured value,  $\hat{y}_i$  is the value predicted by a 3-nearest neighbour algorithm (this algorithm is explained further in the text) and  $N$  is the number of examples. In order to overcome the SVM performance dependence on gamma value [9], a grid search of  $2^{\{-15; -11; -7; -3; 1\}}$  was adopted to optimize the kernel parameter  $\gamma$ , following the same procedure as adopted for ANN, avoiding this way the possible oversmooth problems and conversely, overfitting.

A decision tree (DT) can be used either for classification or for regression [3, 8]. What differs is the type of the target variable, which can be quantitative (regression), respectively, qualitative (classification), in which case the prediction corresponds to a numerical value, respectively, to a class. It is a direct and acyclic flow chart that is graphically

represented by a binary tree. The top node, or root, represents the whole data that is to be predicted ( $T$ ). The other nodes are called the internal nodes. Each internal node is associated with a single predictor and represents a split of the data into subsets. The split is determined by a test or rule on that predictor. Each of the end nodes of the tree, or leaves, contains the prediction that is a numerical value or a class.

The building of the tree is an iterative process: from the root, a rule is defined using a predictor, which splits the dataset into two subsets, expressing a simple and conditional logic. The first node is thus created. On each of the two subsets, a prediction can be made, by averaging the observations contained in the subset. This prediction can thus be compared to each actual value, which permits to build the sum of squared error (SSE). This SSE is the cost function that the building of the node aims at minimizing. This operation is repeated on each of the two subsets, each predictor being possibly used several times. This constitutes the growing of the tree, which stops when a stopping criterion is reached. Most commonly, the stopping criteria are defined based on the number of observations contained in the node. If splitting the subset leads to a number of observations in the subsequent sets lower than a certain value, then the split is not accepted and the node is kept as a leaf node. It can also be based on a threshold value that the SSE on the subset has to respect. In this work, the number of observations contained in the node was adopted as a stopping criteria, which was set as 20 as coded in the R part package [38].

This is known as the CART algorithm, the acronym for classification and regression trees that was used in the present work to carry on the growing. It is one of the most popular algorithms used for inducing decision trees. It grows only binary trees (i.e., trees where only two branches can attach to a single root or node) so, despite its high flexibility, it can sometimes be unreliable and computationally slow. Once the growing is finished thanks to this fully automated process, the resulting tree may be overstructured (i.e., it contains too many nodes) and thus inefficient, because too sensitive to noise, outliers, and it has lost its generalization capacity. To avoid this overfit of the data, the pruning is carried on. Tree pruning attempts to simplify the tree by identifying and removing branches with the goal of improving the accuracy of the prediction. In this work, the pruning of the tree was controlled by defining a complexity parameter (cp), which was set to 0.01. This means that any split that does not decrease the overall lack of fit by a factor of cp is not attempted. In other words, the cp stops the tree from growing, though one can think of that as a sort of pre-emptive pruning.

The greatest benefits of decision trees approach are that they are easy to understand and interpret. They are

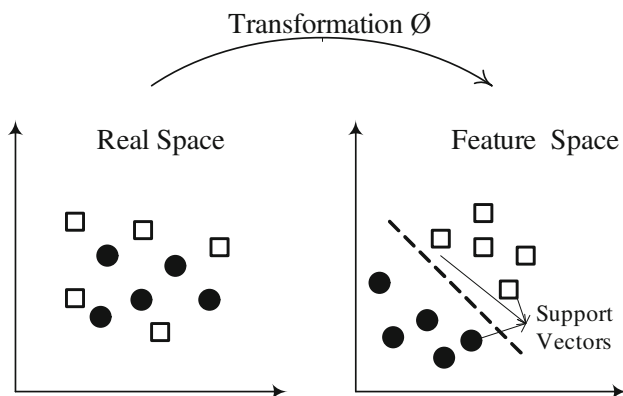


Fig. 4 SVM transformation. Adapted from Cortez [14]



considered a “white box” model, i.e., the induced rules are clear and easy to explain as they use a simple conditional logic. The main drawback is that they get harder to manage as the complexity of data increases leading to a higher number of branches in the tree.

In order to estimate the output  $\hat{Y}(x)$  corresponding to the input  $x$ , the nearest neighbour methods consist in taking into account the observations in the training set whose corresponding inputs  $x_i$  are the closest to  $x$  [17]. Specifically, the  $k$ -nearest neighbour (kNN) prediction for  $x$  is defined as follows:  $\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$ , where  $N_k(x)$  is the neighbourhood of  $x$  defined by the  $k$  closest points  $x_i$  in the training sample. The response is thus an average of the  $k$  selected observations. Closeness implies a metric, which in this work was set to the Minkowski distance [1]. The kNN method requires selection of  $k$ , the number of neighbours, which has a strong effect on the kNN classifier obtained. In this work a grid search of  $\{1; 3; 5; 7; 9\}$  (only considering training data as previously explained for ANN) was adopted during the learning phase to find the best  $k$  value.

Random forest (RF) [7, 15] is a popular ensemble technique, which uses decision trees (DT, as described previously) as building blocks to construct more powerful prediction models. More precisely, this method aims at overcoming the main weakness of a single DT, which is its high variance (its result highly depends on the choice of the learning set), by averaging several decorrelated trees. The idea is to generate  $T$  decision trees by training them on  $T$  different so-called *bootstrapped* training samples. By *bootstrapped*, it is meant that each of the samples contains as many observations as the whole learning set, but those observations are randomly selected among the whole set, the sampling being performed with replacement. It means that the same observation can occur more than once in the bootstrap data set. The  $T$  datasets thus necessarily overlap. The specificity of RF is that when building these DT, each time a split is considered (and a rule is determined), only a sample of predictors is chosen as split candidates among the complete set of predictors. Typically, if  $p$  is the total number of predictors,  $\sqrt{p}$  predictors are considered. Doing so reduces the influence one strong predictor could have on the building of the  $T$  trees, and thus reduces the correlation between them. Once the trees are built, the global prediction is obtained by averaging the predictions of the  $T$  trees. In that way, a strong model is generated, that is less likely to overfit and which balances the bias-variance trade-off. Also for RF, a grid search of  $\{1; 2; 3; 4; 5\}$  was adopted (following the same procedure adopted for ANN) to define  $M_{try}$  (number of variables randomly sampled as candidates at each split) parameter.

In MR, several independent variables are linearly combined to predict the dependent (output) variable [17]. Due

to its additive nature, this model is easy to interpret and is widely used in regression tasks. However, one of its main limitations is its inefficiency at modelling problems of a nonlinear nature. MR was essentially used in this study as a baseline comparison.

As model inputs, a set of four variables were considered to predict the piezometric level (PL): time ( $t$ , days), relative trough ( $Z$ ),  $\cos(S)$  and  $\sin(S)$ . A dataset comprising 623 records monitored on the studied dam was used for models training/validation.

All experiments were conducted using the R statistical environment [38] and supported through the *rminer* package [14], which facilitates the implementation of several DM algorithms, as well as different validation approaches such as cross-validation.

### 3.3 Models evaluation

For models evaluation and comparison, three metrics commonly used in regression problems were used: mean absolute error (MAE), root-mean-square error (RMSE) and the squared correlation coefficient ( $R^2$ ). These metrics are computed based on the difference between the observed and predicted values (the errors). Typically, the lower the error, the better the predictive model, with zero corresponding to the highest model performance. However, while low values of MAE and RMSE should be interpreted as indicating high model predictive capacity,  $R^2$  should be as close as possible to one. The main difference between the MAE and RMSE is that the latter is more sensitive to extreme values because it uses the square of the distance between the real and predicted values. When compared to MAE, RMSE penalizes more heavily a model that produces high errors in a few cases.

Additionally, in order to ease the comparison of different regression models, the regression error characteristic (REC) curve proposed by Bi and Bennett [5] was built, which plots the error tolerance on the  $x$ -axis versus the percentage of points predicted within the tolerance on the  $y$ -axis. The error tolerance is the difference between observed and predicted values (residuals).

The models generalization performance was accessed by five runs under a cross-validation ( $k$ -fold = 3) approach, where the data ( $P$ ) are randomly sampled into  $k$  mutually exclusive subsets ( $P_1, P_2, \dots, P_k$ ), with the same length [17]. Training and testing is performed  $k$  times and the overall error of the model is taken as the average of the errors obtained in each iteration. Under this scheme, all of the data are used for training and testing. Yet, this method requires approximately  $k$  (the number of subsets) times more computation, because  $k$  models must be fitted. In particularly, each modelling setup was trained  $3 \times 5 = 15$  times. Also, the three prediction metrics are always

computed on test unseen data (as provided by the threefold validation procedure).

In addition to the model accuracy, its interpretability is also of high importance, especially from an engineer viewpoint. However, and in particular SVM and ANN algorithms, which rely on complex statistical analysis, are frequently referred to as “black boxes” due to their high complexity. To overcome this drawback of data-driven models, Cortez and Embrechts [13] proposed a novel visualization approach based on sensitivity analysis (SA), which was used in this work. SA is a simple method that is applied after the training phase and measures the model responses when a given input is changed, allowing the quantification of the relative importance of each attribute as well as its average effect on the target variable.

In particular, the global sensitivity analysis (GSA) method [13] was applied, which is able to detect interactions among input variables. This is achieved by performing a simultaneous variation of  $F$  inputs. Each input is varied through its range with  $L$  levels and the remaining inputs fixed to a given baseline value. In this work was performed adopted the average input variable value as a baseline and set  $L = 12$ , which allows an interesting detail level under a reasonable amount of computational effort.

With the sensitivity response of the GSA, two important visualization techniques were computed. First the input importance barplot was built, which shows the relative influence ( $R_a$ ) of each input variable in the model. To measure this effect, the gradient metric ( $g_a$ ) for all inputs was calculated. After that, the relative influence was computed.

$$R_a = g_a / \sum_{i=1}^I g_i \cdot 100 (\%)$$

$$\text{where, } g_a = \sum_{j=2}^L |\hat{y}_{a,j} - \hat{y}_{a,j-1}| / (L - 1) \quad (1)$$

where  $a$  denotes the input variable under analysis,  $I$  is the number of input variables and  $\hat{y}_{a,j}$  is the sensitivity response for  $x_{a,j}$ .

Second, in order to analyse the average impact of a given input in the fitted model, the variable effect characteristic (VEC) curve was used. For a given input variable, the VEC curve plots the attribute  $L$  level values ( $x$ -axis) versus the SA responses ( $y$ -axis). Between two consecutive  $x_{a,j}$  values, the VEC plot performs a linear interpolation. To enhance the visualization analysis, several VEC curves can be plotted in the same graph. In such case, the  $x$ -axis is scaled (e.g., within  $[0, 1]$ ) for all  $x_a$  values.

In order to show the gains of performance when using the DM approach when compared with the current practice, the HST model was also applied to the dataset, because it is the most popular in dam engineering. The above-

mentioned evaluation criteria were thus also applied to this multilinear model so as to compare it with the DM algorithms.

## 4 Results analysis and discussion

The average hyperparameters and fitting time values (and respective 95% confidence intervals according to Student's  $t$  distribution) of all DM models are shown in Table 2. Concerning the running time, MR and DT are extremely fast to fit. The slowest one is the RF, which takes an average of 16 s over the five runs. ANN and SVM, that achieved the best performance, took, on average, around 3.7 and 1.2 s over the five runs, respectively. These computational times are related to the time that each algorithm took to fit the training data. In the future, when the proposed models (namely the ANN model) are applied to predict new cases, the time required is very close to zero (the computation is almost instantaneous).

Table 3 compares the performance of the six DM algorithms in PL prediction based on MAE, RMSE and  $R^2$  metrics (mean value and respective 95% level confidence intervals according to a  $t$ -student distribution). Apart from MR and DT, all algorithms present a very good response in PL prediction, with a  $R^2$  very close to one. The highest performance in PL prediction was achieved by the ANN model, with an  $R^2 = 0.9912$ , very closely followed by the SVM ( $R^2 = 0.9859$ ).

Figure 5 compares REC curves of all models confirming the poor performances of MR, DT and HST. Figure 5 also highlights the high accuracy of the ANN and the SVM, showing that both models are able to predict around 96% of all records with an absolute deviation lower than 5 m. Even for a tighter tolerance, such as an absolute deviation around 2.5 m, ANN presents an accuracy higher than 85%. For very high tolerances (lower than 2.5 m), the RF and kNN perform slightly better when compared to SVM.

**Table 2** Hyperparameters and computation time of each DM model

Model	Hyperparameters	Time (s)
ANN	$H = 8 \pm 1$	$3.688 \pm 0.085$
SVM	$\gamma = 0.5 \pm 0.01, \varepsilon = 0.1 \pm 0.0$	$1.182 \pm 0.144$
DT	–	$0.072 \pm 0.014$
KNN	$k = 3 \pm 1$	$0.322 \pm 0.016$
RF	$M_{\text{try}} = 5 \pm 2$	$16.074 \pm 0.725$
MR	–	$0.056 \pm 0.007$

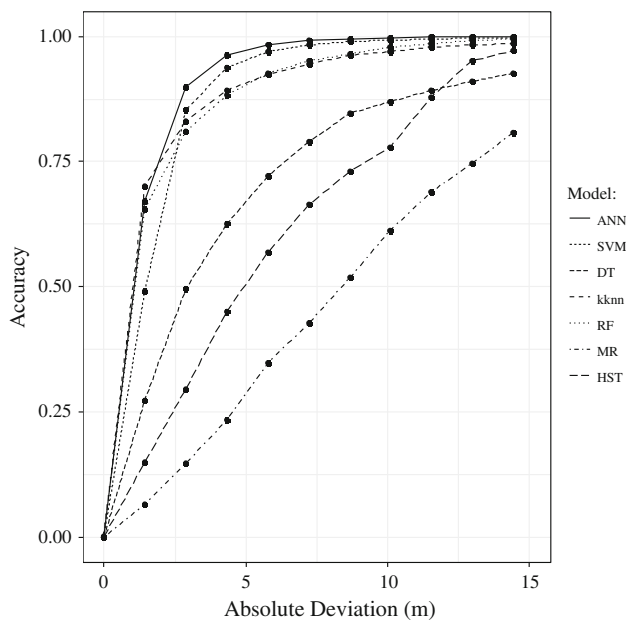
The uncertainty derives from the five repetitions (five runs) applied for models generalization assessment

**Table 3** Models performance comparison based on MAE, RMSE and  $R^2$  metrics

Model	MAE	RMSE	$R^2$
ANN	<b>1.35 ± 0.09</b>	<b>1.97 ± 0.13</b>	<b>0.9912 ± 0.0012</b>
SVM	<b>1.80 ± 0.09</b>	<b>2.56 ± 0.18</b>	<b>0.9859 ± 0.0019</b>
DT	4.89 ± 0.32	7.35 ± 0.56	0.8782 ± 0.0181
KNN	1.77 ± 0.12	3.62 ± 0.33	0.9708 ± 0.0054
RF	1.81 ± 0.07	3.13 ± 0.11	0.9781 ± 0.0016
MR	9.20 ± 0.09	11.20 ± 0.31	0.7174 ± 0.0143
HST	5.98	7.51	0.8726

The uncertainty derives from the five repetitions (five runs) applied for models generalization assessment

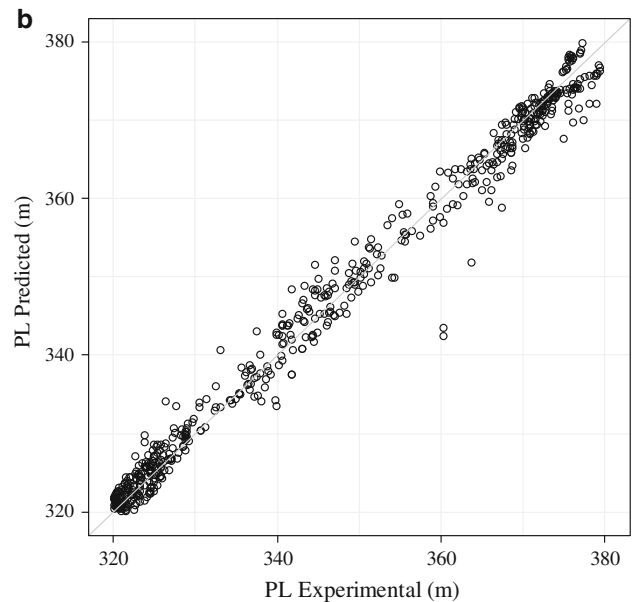
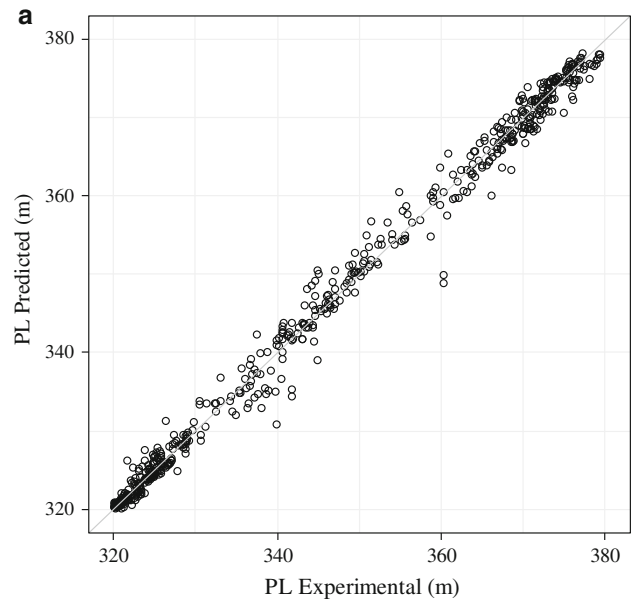
Best values are indicated in bold



**Fig. 5** Models performance comparison based on REC curves [relation between tolerance (residuals (m)) and cumulative correct predictions within the error tolerance]

Figure 6 depicts the relation between observed and predicted PL values (scatterplot) according to ANN (Fig. 6a) and SVM (Fig. 6b) models, showing once again a very interesting fit (all points are very close to the diagonal line).

On Fig. 7 is plotted the relative importance of each input according to the six DM models. Note that since those inputs were determined in order to represent the three particular external loads (hydrostatic, thermal, and temporal) that affect the behaviour of the dam, which is why  $\cos S$  and  $\sin S$  have to be interpreted jointly. From its analysis, there is no doubt that  $Z$  is the most relevant variable in PL prediction, which is confirmed by the six algorithms. Taking ANN model as reference,  $Z$  has a

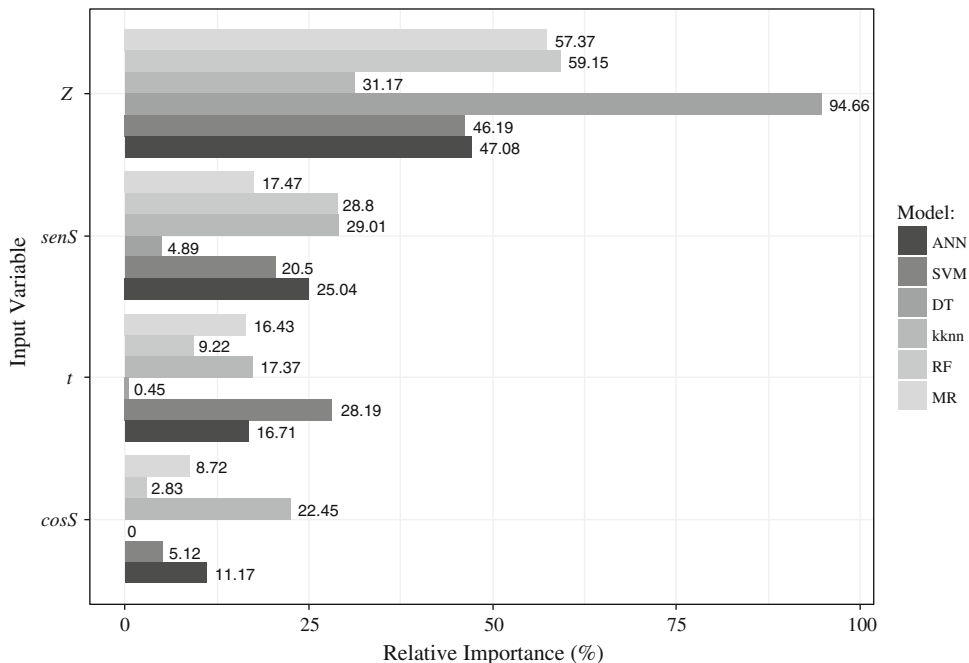


**Fig. 6** Scatterplot of: **a** ANN; **b** SVM

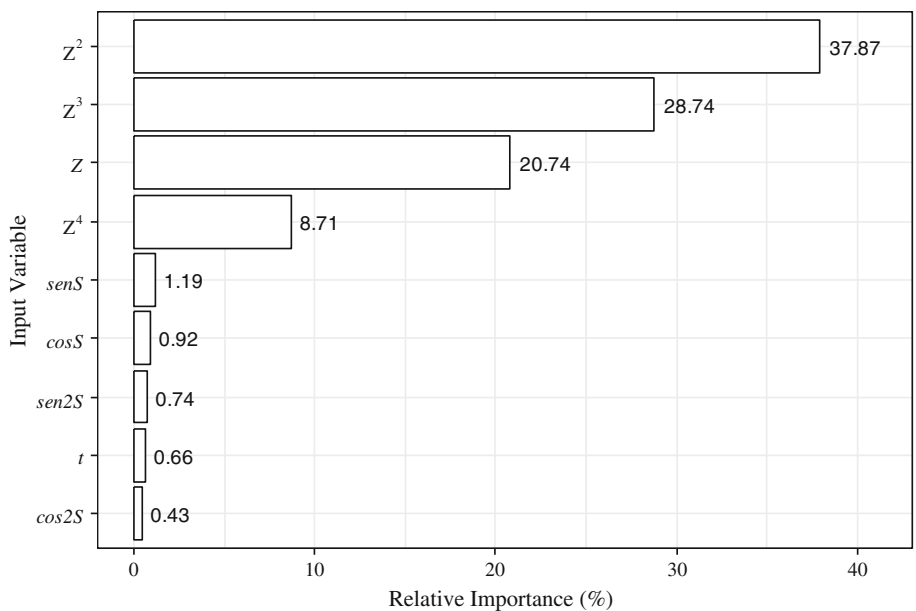
relative influence around 50%, followed by  $\sin S$  and  $\cos S$  with around 36%. Time is the least influencing variable. The same classification can be drawn from the relative importance analysis conducted with HST model (Fig. 8): the four most relevant variables, representing more than 96%, are  $Z^2$ ,  $Z^3$ ,  $Z$  and  $Z^1$ . The sensitivity of the PL to the thermal load represents less than 4% of the total, and time comes last.

This plot (Fig. 7) permits to classify the inputs relatively compared to each other; however, it has to be considered with caution because it does not in any case assess to what absolute extent they explain the PL. The purpose is not to

**Fig. 7** Relative importance comparison of each input variable according to the six DM models



**Fig. 8** Relative importance of each input variable according to HST model



select or leave inputs out of the analysis. The main interest of this analysis is that it shows that the studied PL are very sensitive to the water level variations. Since those piezometric levels are an image of the aperture of the rock–concrete interface, it means that the size of the aperture is very much correlated with the annual variations of the water level. This is in accordance with engineering knowledge and gives even more credit to the model. From an engineer point of view, this also shows that a way to

control the aperture and limit its expansion is to adapt the water level.

In terms of physical behaviour, this relative importance analysis shows that the hydrostatic load plays a dominating role in determining the state of stress of the dam. The thermal load comes next, and although time is the least significant influence quantity, it does not mean that it should be removed from a diagnostic analysis. Indeed, the evolution induced by time might be slow, so it might not

show in a spectacular way compared to the evolutions induced by the other influence quantities. However, it might represent some slow irreversible evolutions of the dam, which on the long term might alter the global integrity of the structure. The propagation of the aperture of the rock–concrete interface corresponds to such a slow but significant phenomenon.

Although the effects of the thermal and the hydrostatic loads are quantitatively the most significant effects, they correspond to elastic evolutions, which imply that it does not induce permanent deformation of the structure. For instance, if a high water level implies a significant movement downstream, simply lowering the water level will permit to have the dam come back to a safe position. Consequently, the effects of the hydrostatic and the seasonal loads in usual ranges are not an issue of too great concern. This is not the case for the irreversible evolutions that appear as time passes. In the context of dam safety, those temporal evolutions have to be identified and explained as early as possible, in order to conduct maintenance operations if needed.

Since ANN model achieved one of the highest performances, the analysis was pushed further for that algorithm. In order to get a better understanding of the PL and of the global behaviour of the structure, the effect of each load on the PL prediction was measured based on a SA [13]. In order to draw some physical conclusions on each model, the SA was performed reasoning in terms of load and not in terms of input. This means that to study the effect of the hydrostatic load, respectively the temporal load, two respective 1-D SAs were performed, having only  $Z$ , respectively  $t$ , vary through its range. In order to study the impact of the thermal load, a 2-D SA was used, having both  $\cos S$  and  $\sin S$  vary simultaneously.

A similar SA was also performed for the HST predictions so as to compare the performances of the ANN model to the reference model. Likewise, for this model, the sensitivity of the PL to the hydrostatic load was assessed having  $Z$ ,  $Z^2$ ,  $Z^3$  and  $Z^4$  vary simultaneously, the sensitivity to the thermal load was assessed having  $\cos S$ ,  $\sin S$ ,  $\cos 2S$  and  $\sin 2S$  vary simultaneously and for the temporal load, a mere 1-D SA was performed.

Accordingly, Fig. 9 overlaps the VEC curves of the temporal load ( $t$ -variable), the hydrostatic load ( $Z$ -variable) and  $S$  and the thermal load ( $S$ -variables).  $xx$  axis is scaled to accommodate all loads. Similarly, Fig. 10 shows the VEC curves corresponding to each load, based on the HST predictions.

Focusing first on the influence of the  $Z$ -variable, what is noticeable on the ANN predictions is that its effect is nearly non-existent when it varies between 0.5 and 1. However, when decreasing  $Z$  below 0.5, PL start to rise, and  $Z$  stands out as the most influencing quantity. Going

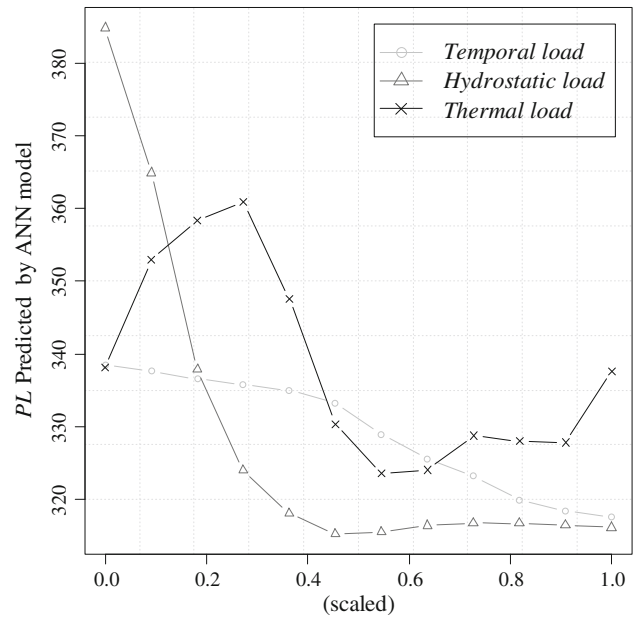


Fig. 9 VEC curves of the temporal, hydrostatic and thermal loads according to the ANN model in PL prediction, based on a SA

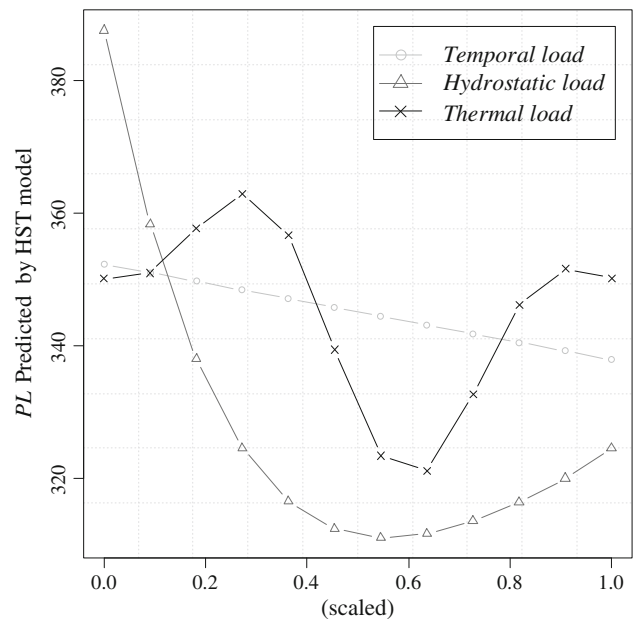


Fig. 10 VEC curves of the temporal, hydrostatic and thermal loads according to the HST model in PL prediction, based on a SA

back to the definition of the relative trough  $Z$ , one notices that  $Z$  is maximal when the water level  $h$  is minimal and vice versa. Thus, Fig. 9 shows that when the water level in the reservoir increases, the PL at the interface increase as well, which is perfectly consistent with the state of the art. What is particularly interesting here is that the rise of the PL occurs only after a threshold is reached (0.5), which can be linked to the state of aperture of the rock–concrete



interface. The interpretation of this nonlinearity is that this threshold water level corresponds to the moment at which the interface starts to open, leading to the hydrostatic load being transmitted to the foundations, and subsequently the rise of the PL. Once the aperture is open, the higher the water level the higher the PL.

This nonlinear feature that characterizes the evolution of the aperture of the rock–concrete interface is what made indispensable the use of models that are more advanced than additive models, such as HST. Indeed, looking at the Z VEC curve obtained using HST, a similar minimum is reached at around 0.5 but as significant difference is that the PL tend to rise when Z rises from 0.5 to 1, meaning that HST predicts that the PL rise when the water level decreases. Such a trend is due to the use of a polynomial law to model the effect of the hydrostatic load, which makes it impossible to model threshold effects. Consequently, the linearity of the reference HST model is not suitable to describe such a phenomenon as the aperture of the rock–concrete interface, while nonlinear models such as ANN are much more adequate.

Second, the curve corresponding to the influence of the thermal load traduces the sensitivity of the PL to the variations of the temperature of the body of the dam over the year, which follows a seasonal variation. For the ANN model, the corresponding curve has a quasi-sinusoidal shape, with the maximum being reached approximately at the first quarter of the year, which is April, and the minimum being reached in summer. The minimum is not perfectly clear though, because one point seems to distort the curve and draw a second maximum, though a smooth sinusoidal shape is expected. Observing the maximum predicted PL during the cold season is consistent with the engineering knowledge of how the thermal load influences the structure. Indeed, the thermal sensitivity of concrete induces a contraction of concrete which leads to a global downstream movement of the dam. This movement is conducive to the expansion of the rock–concrete aperture, which eventually results in the rise of the PL. However, because of the thermal inertia of concrete, there is a delay between the air temperature minimum, reached in average between January and February in the region where the dam is situated, and the maximum temperature of the body of the dam (including the foundations). Conversely, the minimum predicted PLs are observed during the hot season, which coincides with the moment of the year when the thermal state lets concrete expand and causes the upstream movement of the arch, minimizing the strain on the rock–concrete interface. Consequently, the aperture of the contact tends to close and eventually the PL decrease. Thus this sensitivity analysis confirms the validity of the ANN which describes accurately the response of the dam to the thermal load. Looking at the corresponding curve obtained

with HST model, the global shape is also sinusoidal, but two maxima are observed. This distortion is even more pronounced than that corresponding to the ANN. The reason for that may be that, in the case of the aperture of the contact, all three loads act simultaneously but not in an additive way; however, HST models the effect of each load by summing the irreversible function, the hydrostatic function and the seasonal function. Thus, it artificially separates the effect of each load on the PL. Consequently, when separating those effects, the model shares out the influences, and some effects eventually compensate others, which might result in the hydrostatic effect rising between 0.5 and 1, and the seasonal effect decreasing between approximately 0.9 and 1. Conversely, because of its flexibility, the ANN is able to produce some crossed effect between the different loads; however, the limit of the SA that is performed here lies in the fact that the non-varying inputs are set to their average value, which is not representative of all real load cases.

Finally, the curve corresponding to the time variable is also of great interest, because it shows that the PL decrease as  $t$  increases. In terms of structural behaviour, it means that PL decreases when time passes, and since the analysed PL are directly linked to the aperture of the interface, this curve shows that the aperture is gradually closing on the period of analysis. From an engineer point of view, the closing of the interface is synonymous with an improvement of the behaviour of the whole structure, and thus an enhancement of its safety. The temporal VEC curve obtained with HST shows a similar decreasing trend, but since the temporal function is a one order linear law, the slope is necessarily constant, which does not allow to distinguish the variations of kinetic that are visible with the ANN predictions.

## 5 Conclusion

This work proposed to challenge six data mining (DM) techniques in order to determine which of them was the most suitable to serve dam monitoring purposes. More particularly, the comparison was based on the analysis of piezometric measurements that were recorded at the rock–concrete interface of an arch dam, in order to get a better understanding of the phenomenon of the aperture of the interface. The study of this nonlinear phenomenon included an HST analysis, which is the historical multilinear regression model that is classically used by civil engineers in dam monitoring. The six DM algorithms were fed with the same four basic inputs: time, the sine and cosine of the season, and the scale water level. In contrast, HST has nine inputs, all derived from the four previous ones. This comparison has highlighted the need for such advanced

techniques as DM techniques to deal with the nonlinear features that are at stake.

Excluding the multiregression (MR) model, all the remaining algorithms outperformed HST, and ANN stood out as the most performing algorithm in terms of prediction, closely followed by SVM. Consequently, the analysis was pushed further for the ANN model, taking advantage of its interpretability.

In order to draw conclusions on the behaviour of the structure, a sensitivity analysis (SA) was performed both for ANN and HST models, based on relative importance plots and Variable Effect Characteristic (VEC) curves. This SA was carried out in order to show the influence of the three external physical loads that affect the behaviour of a dam, namely the hydrostatic load, the thermal load and the influence of time. This means that instead of studying the influence of the model's inputs individually, some groupings of those variables were considered, so as to be representative of a physical reality. While HST was unable to detect the coupled influence of the loads on the piezometric levels, it appeared that ANN could adapt better, and the observed effects were in accordance with the engineering knowledge. This work thus shows that ANN can describe more adequately the effect of the influencing loads. This leads to a better understanding of how the aperture of the rock–concrete interface evolves, which is a significant issue for dam safety.

This work also detailed the performances of a multiregression (MR) model for which the four basic inputs were provided, with HST, for which those four inputs were supplemented by derived forms of those four inputs. Those additional inputs permit to define a polynomial law to describe the hydrostatic effect, and supplement the sine and cosine of the season with their second-order harmonic, in order to be more exhaustive in the description of the annual thermal waves that impact the dam. Those more complete hydrostatic and thermal laws were determined empirically. What is interesting to see when comparing HST and this MR is that adding pertinent input variables greatly improves the predictions as shown on Fig. 5 and Table 3. Noticeably, DM techniques are flexible enough to build thresholds and combine the inputs so as to compute correct effects. Thus, suppose that one variable is approached for having an influence on a given phenomenon, DM algorithms can be used to add “blindly” this new input. The work of the engineer will then be to interpret the outputs of those algorithms to define how this input impacts the studied phenomenon. Subsequently, this understanding of the impact of this new variable can help the engineer determine new pertinent inputs to be used as inputs of simpler models, which can bring major improvement to the analysis, and ease of use. The advantage of this type of

approach is highlighted by this comparison between HST and the MR model.

DM techniques can indubitably provide great improvement to the dam monitoring profession, but interpreting them falls within the competence of experienced engineers. What is more, because of their complexity, advanced algorithms are often more difficult to tune than simpler models, and their flexibility requires that they should be specifically adapted to the studied phenomenon. Consequently, those promising techniques should be associated with the engineer experience and will permit a better understanding of the structure behaviour, especially when safety is at stake.

**Acknowledgements** The authors thank the ANRT CIFRE for its Grant (No. 0902/2016) that partly supported this work.

### Compliance with ethical standards

**Conflict of interest** There is no conflict of interest between the authors.

### References

1. Aggarwal C, Reddy C (eds) (2013) Data clustering: algorithms and applications. CRC Press, Boca Raton
2. Balcilar M, Demirkaya S (2012) The contribution of soft computing techniques for the interpretation of dam deformation. FIG Working Week, Rome
3. Berry M, Linoff G (2000) Mastering data mining: the art and science of customer relationships management. Wiley, New York
4. Bhattacharya S, Roy S, Chowdhury S (2018) A neural network-based intelligent cognitive state recognizer for confidence-based e-learning system. *Neural Comput Appl* 29(1):205–219
5. Bi J, Bennett K (2003) Regression error characteristic curves. In: Proceedings of the twentieth international conference on machine learning. AAAI Press, Washington, pp 43–50
6. Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Oxford
7. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
8. Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
9. Boukouvalas A, Cornford D, Stehlik M (2014) Optimal design for correlated processes with input-dependent noise. *Comput Stat Data Anal* 71:1088–1102
10. Carrère A, Colson M, Goguel B, Noret C (2000) Modelling: a means of assisting interpretation of readings. In: XXth international congress on large dams, vol III, Beijing, pp 1005–1037
11. Cherkassky V, Ma Y (2004) Practical selection of SVM parameters and noise estimation for svm regression. *Neural Netw* 17(1):113–126
12. Cortes C, Vapnik V (1995) Support vector networks. *Mach Learn* 20(3):273–297
13. Cortez P, Embrechts M (2013) Using sensitivity analysis and visualization techniques to open black box data mining models. *Inf Sci* 225:1–17
14. Cortez P (2010) Data Mining with neural networks and support vector machines using the R/rminer tool. In: proceedings of advances in Data Mining—applications and theoretical aspects

- 10th industrial conference on data mining (ICDM 2010). Lecture notes in artificial intelligence, vol 6171, pp 572–583
15. Genuer R, Poggi JM, Tuleau-Malot C, Villa-Vialaneix N (2017) Random forests for big data. *Big Data Res* 9:28–46
  16. Guedes Q, Coelho P (1985) Statistical behaviour model of dams. 15th ICOLD congress, pp 319–334
  17. Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference, and prediction*. Springer, New York
  18. ICOLD (2001) *Proceedings of the sixth ICOLD benchmark workshop on numerical analysis of dams*, Salzburg, Austria
  19. Kang F, Li J-S, Wang Y, Li J (2017) Extreme learning machine based surrogate model for analyzing system reliability of soil slopes. *Eur J Environ Civ Eng* 21(11):1341–1362
  20. Kenig S, Ben-David A, Omer M, Sadeh A (2001) Control of properties in injection molding by neural networks. *Eng Appl Artif Intell* 14(6):819–823
  21. le Delliou P (2003) *Les barrages: conception et maintenance*, 2nd edn. In: de Lyon PU (ed)
  22. Léger P, Leclerc M (2007) Hydrostatic, temperature, time-displacement model for concrete dams. *J Eng Mech* 133(3):267–277
  23. Mata J (2011) Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models. *Eng Struct* 33:903–910
  24. Mendes de Vasconcelos Braga Farinha ML (2010) Hydromechanical behaviour of concrete dam foundations. In *Situ tests and numerical modelling*. Universidade Técnica de Lisboa Instituto Superior Técnico
  25. Murthy VM, Kumar A, Sinha PK (2018) Prediction of throw in bench blasting using neural networks: an approach. *Neural Comput Appl* 29(1):143–156
  26. Penot I, Daumas B, Fabre J-P (2005) Monitoring behaviour. *Water Power Dam Constr* 57(12):24–27
  27. Ranković V, Grujović N, Divac D, Milivojević N (2014) Development of support vector regression identification model for prediction of dam structural behaviour. *Struct Saf* 48:33–39
  28. Ranković V, Grujović N, Divac D, Milivojević N, Novaković A (2012) Modelling of dam behaviour based on neuro-fuzzy identification. *Eng Struct* 35:107–113
  29. Ranković V, Novaković A, Grujović N, Divac D, Milivojević N (2014) Predicting piezometric water level in dams via artificial neural networks. *Neural Comput Appl* 24:1115–1121
  30. Rumelhart DE, Hint GE, Williams RJ (1985) Learning internal representations by error propagation. *Parallel Distrib Process Explor Microstruct Cogn* 1:318–362
  31. Safarzaghean Gilan S, Bahrami Jovein H, Ramezaniyanpour A (2012) Hybrid support vector regression-particle swarm optimization for prediction of compressive strength and rcpt of concretes containing metakaolin. *Constr Build Mater* 34:321–329
  32. Salazar F, Morán R, Toledo MÁ, Oñate E (2015) Data-based models for the prediction of dam behaviour: a review and some methodological considerations. *Arch Comput Methods Eng* 24:1–21
  33. Salazar F, Oñate E, Toledo MA (2017) A machine learning based methodology for anomaly detection in dam behaviour. CIMNE, Barcelona
  34. Salazar F, Toledo MA, Oñate E, Morán R (2015) An empirical comparison of machine learning techniques for dam behaviour modelling. *Struct Saf* 56:9–17
  35. Santillán D, Fraile-Ardanuy J, Toledo MA (2013) Dam seepage analysis based on artificial neural networks: the hysteresis phenomenon. In: *Proceedings of the international joint conference on neural networks*. IEEE, pp 1–8
  36. Simon A, Royer M, Mauris F, Fabre J-P (2013) Analysis and interpretation of dam measurements using artificial neural networks. In: 9th ICOLD European club symposium, Venice
  37. Smola A, Schölkopf B (2004) A tutorial on support vector regression. *Stat Comput* 14(3):199–222
  38. Team R (2009) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. <http://www.r-project.org/>
  39. Tinoco J, Gomes Correia A, Cortez P, Toll DG (2017) Stability condition identification of rock and soil cutting slopes based on soft computing. *J Comput Civ Eng* 32(2):04017088
  40. Venables W, Ripley B (2003) *Modern applied statistics with S*, 2nd edn. Springer, Heidelberg
  41. Willm G, Beaujoint N (1967) Les méthodes de surveillance des barrages au service de la production hydraulique d'Electricité de France, problèmes anciens et solutions nouvelles. In: IXth international congress on large dams, Istanbul, pp 529–550
  42. Young CC, Liu WC, Chung CE (2015) Genetic algorithm and fuzzy neural networks combined with the hydrological modeling system for forecasting watershed runoff discharge. *Neural Comput Appl* 26(7):1631–1643

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.