An empirical comparison of machine learning techniques for dam behaviour modelling

F. Salazar[1], M.A. Toledo[2], E. Oñate[1,3], R. Morán[2]

Abstract

Predictive models are essential in dam safety assessment. Both deterministic and statistical models applied in the day-to-day practice have demonstrated to be useful, although they show relevant limitations at the same time. On another note, powerful learning algorithms have been developed in the field of machine learning (ML), which have been applied to solve practical problems. The work aims at testing the prediction capability of some state-of-the-art algorithms to model dam behaviour, in terms of displacements and leakage. Models based on random forests (RF), boosted regression trees (BRT), neural networks (NN), support vector machines (SVM) and multivariate adaptive regression splines (MARS) are fitted to predict 14 target variables. Prediction accuracy is compared with the conventional statistical model, which shows poorer performance on average. BRT models stand out as the most accurate overall, followed by NN and RF. It was also verified that the model fit can be improved by removing the records of the first years of dam functioning from the training set.

Keywords: Dam monitoring, Dam safety, Machine learning, Boosted regression trees, Neural networks, Random forests, MARS, Support vector machines, Leakage flow

Email addresses: fsalazar@cimne.upc.edu (F. Salazar), matoledo@caminos.upm.es (M.A. Toledo), onate@cimne.upc.edu (E. Oñate), rmoran@caminos.upm.es (R. Morán)
URL: www.cimne.com (F. Salazar)
[1]International Center for Numerical Methods in Engineering (CIMNE). Campus Norte UPC. Gran Capitán s/n. 08034. Barcelona, Spain
[2]Technical University of Madrid (UPM). Civil Engineering Department: Hydraulics and Energy. Profesor Aranguren s/n, 28040, Madrid, Spain
[3]Universitat Politècnica de Catalunya (UPC). Barcelona, Spain

1. Introduction and background

Dam safety assessment is a complex task due to the uniqueness of each of such structures and their foundations. It is commonly based on three main pillars: visual inspection, engineering knowledge and a behaviour model. The actual response of the dam is compared with the predictions of the model, with the aim of detecting anomalies and preventing failures. Current predictive methods can be classified as follows [1]:

- Deterministic: typically based on the finite element method (FEM), these methods calculate the dam response on the basis of the physical governing laws.
- Statistical: exclusively based on dam monitoring data.
- Hybrid: deterministic models which parameters have been adjusted to fit the observed data.
- Mixed: comprised by a deterministic model to predict the dam response to hydrostatic pressure, and a statistical one to consider deformation due to thermal effects.

It is difficult to predict dam behaviour with high accuracy. Numerical models based on the FEM provide useful estimates of dam movements and stresses, but are subject to a significant degree of uncertainty in the characterisation of the materials, especially with respect to the dam foundation. Other assumptions and simplifications have to be made, regarding geometry and boundary conditions. These tools are essential during the initial stages of the life cycle of the structure, provided that there are not enough data available to build data-based predictive models. However, their results are often not accurate enough for a precise assessment of dam safety.

This is more acute when dealing with leakage in concrete dams and their foundations, due to the intrinsic features of the physical process, which is often non-linear [2], and responds to threshold and delayed effects [3], [4]. Numerical

analysis cannot deal with such a phenomenon, because comprehensive information about the location, geometry and permeability of each fracture would be needed. As a result, deterministic models are not used in practice for the prediction of leakage flow in concrete dams [1].

Many of the dams in operation have a large number of monitoring devices, recording the evolution of various indicators such as movements, leakage flow or the pore water pressure, among others. Although there are still many dams with few observed data, there is a clear trend towards the installation of a larger number of devices with higher data acquisition frequency [5]. As a result, there is an increasing amount of information on the dam performance, which makes it interesting to study the ability of machine learning (ML) tools to process them, build behaviour models and extract useful information [6].

The paper assesses the potential of some state-of-the-art ML techniques to build models for the prediction of dam behaviour. The results are compared with the conventional statistical method.

1.1. Statistical models

The most popular data-driven approach for the prediction of dam behaviour is the hydrostatic-seasonal-time (HST) method, characterised by taking into account three effects:

- A reversible effect of the hydrostatic load.
- A reversible seasonal thermal influence of the temperature.
- An irreversible term due to the evolution of the dam response over time.

This assumption is coherent with the observed behaviour of many concrete dams in terms of displacements. However, it has also been applied to other variables, such as uplifts and leakage [3]. Similar schemes have been used for rock-fill dams, although it is acknowledged that the temperature is not relevant, and that the irreversible effect of settlements prevails on the elastic response to hydrostatic load. Furthermore, rainfall may have a strong influence on leakage [3].

The main drawbacks of HST and other methods based on linear regression are the following:

- The functions have to be defined beforehand, and thus may not represent the true behaviour of the structure [3].
- The governing variables are supposed to be independent, although some of them have been proven to be correlated [7].
- They are not well-suited to model non-linear interactions between input variables [2].

1.2. Advanced data analysis in dam monitoring

The examples of application of innovative techniques to improve the results of HST are becoming more frequent in recent years. As an example, Bonelli and Radzicki [8] used an impulse-response function for predicting the pore pressure in the dam body. The method provided accurate results in the test cases, showing the hysteresis effect by which the pore pressure is lower during filling than it should be in a stationary state, and vice versa. Nonetheless, given that it makes a strong assumption on the characteristics of the phenomenon, it is restricted to specific processes.

Li et *al.* [9] proposed a method to improve HST based on cointegration theory. They tested the stationarity of the monitoring data series before fitting a multi-linear regression (MLR) model.

One obvious weakness of linear regression is that it cannot reproduce non-linear relations between variables. This problem is typically overcome by introducing higher order terms of the covariates. Neural networks (NN) constitute a powerful alternative to solve this issue. Their flexibility and capability to adapt to highly complex interactions have made them popular in several fields of engineering, including dam monitoring (see for example [3], [10], [11], and [12]).

However, it should be noted that NN have some drawbacks:

- The result depends on the initialisation of the weights.
- The best network architecture (number of hidden layers and neurons in each layer) is not known beforehand.
- The model is prone to over-fitting.
- The training process may reach a local minimum of the error function.

Several techniques have been developed to overcome these shortcomings, which in general lead to an increase in the computational cost [13]. In spite of this, NN stand out as the most popular ML tool in dam engineering, and the results are promising [3]. Further models have been also applied to dam monitoring, such as ANFIS (adaptive network-based fuzzy inference system) models [14], principal component analysis [6], NARX (nonlinear autoregressive with exogenous input) models [15] or K-nearest neighbours [16]. However, these tools are rarely used in practice, where HST still prevails. Moreover, most of the previous studies are limited to one single variable of specific dams [11], [12]. Hence, the results are not generally applicable.

1.3. Objectives

The study aims to assess the prediction accuracy of some ML tools, most of which have been seldom used in dam engineering. Specifically, the algorithms selected are: random forests (RF), boosted regression trees (BRT), support vector machines (SVM) and multivariate adaptive regression splines (MARS). Both HST and NN were also used for comparison purposes. Similar analyses have been performed in other fields of engineering, such as the prediction of urban water demand [17].

A further singularity of dams is that the early years of operation often correspond to a transient state, non-representative of the quasi-stationary response afterwards [18]. In such a scenario, using those years for training a predictive model would be inadvisable. This might lead to question the optimal size of the training set in achieving the best accuracy. De Sortis [19] ran a sensitivity

analysis and concluded that at least 10 years were needed to obtain acceptable predictions. However, his study was limited to the prediction of the radial displacement in one particular location of a specific dam by using HST. A similar work was performed by Chouinard and Roy [2]. This paper seeks to extend such studies to other learning algorithms and output variables.

2. Case study and variable selection

The data used for the study correspond to La Baells dam. It is a double curvature arch dam, with a height of 102 m, which entered into service in 1976. The monitoring system records the main indicators of the dam performance: displacement, temperature, stress, strain and leakage. The data were provided by the Catalan Water Agency (*Agència Catalana de l'Aigua*, ACA), the dam owner, for research purposes. Among the available records, the study focuses on 14 variables: 10 correspond to displacements measured by pendulums (five radial and five tangential), and four to leakage flow. Several variables of different types were considered in order to obtain more reliable conclusions. Table 1 shows some statistics of the target variables, whereas the location of each monitoring device is depicted in Figure 1.

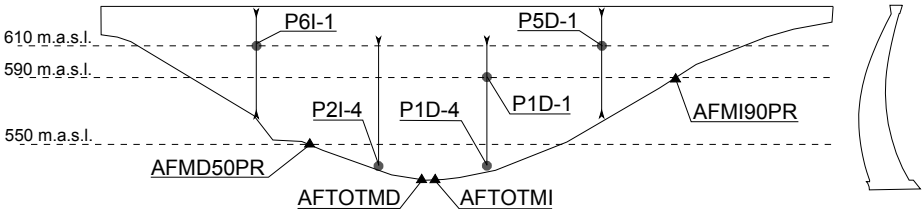

Figure 1: Geometry and location of the monitoring devices in La Baells Dam. Left: view from downstream. Right: highest cross-section.

The data acquisition frequency is of the order of one record per week. The measurement error of the devices is $\pm 0.1mm$ for displacements, and negligible for leakage flows (measured using the volumetric method). The series span from 1979 to 2008. In all cases, approximately 40% of the records (from 1998 to 2008) were left out as the testing set. This is a large proportion compared with

Target	# Observations	Type	units	Mean	Min.	Max.
P1DR1	1,194	Radial displ.	mm	-10.75	-20.6	2.1
P1DR4	1,194	Radial displ.	mm	-9.88	-16.8	0.0
P2IR4	1,191	Radial displ.	mm	-7.77	-17.5	1.3
P5DR1	1,193	Radial displ.	mm	-6.37	-14.8	1.9
P6IR1	1,198	Radial displ.	mm	-9.24	-17.5	0.1
P1DT1	1,194	Tangential displ.	mm	2.36	0.0	3.9
P1DT4	1,194	Tangential displ.	mm	-0.32	-1.5	0.3
P2IT4	1,191	Tangential displ.	mm	-1.56	-2.7	-1.1
P5DT1	1,193	Tangential displ.	mm	-0.09	-1.8	1.6
P6IT1	1,199	Tangential displ.	mm	-2.04	-4.2	0.1
AFMD50PR	1,016	Leakage	l/min	5.05	0.0	27.3
AFMI90PR	994	Leakage	l/min	0.63	0.0	3.0
AFTOTMD	1,064	Leakage	l/min	7.30	0.1	35.8
AFTOTMI	1,014	Leakage	l/min	2.89	0.1	12.4

Table 1: Main statistics of the target variables in the interval considered (1.997-2.008).

previous studies, which typically leave 10-20 % of the available data for testing [14], [12], [16]. A larger test set was selected in order to increase the reliability of the results.

Four different training sets were chosen to fit each model, spanning five, 10, 15 and 18 years of records. In all cases, the training data used correspond to the closest time period to the test set (e.g. periods 1993-1997, 1988-1997, 1983-1997, and 1979-1997, respectively).

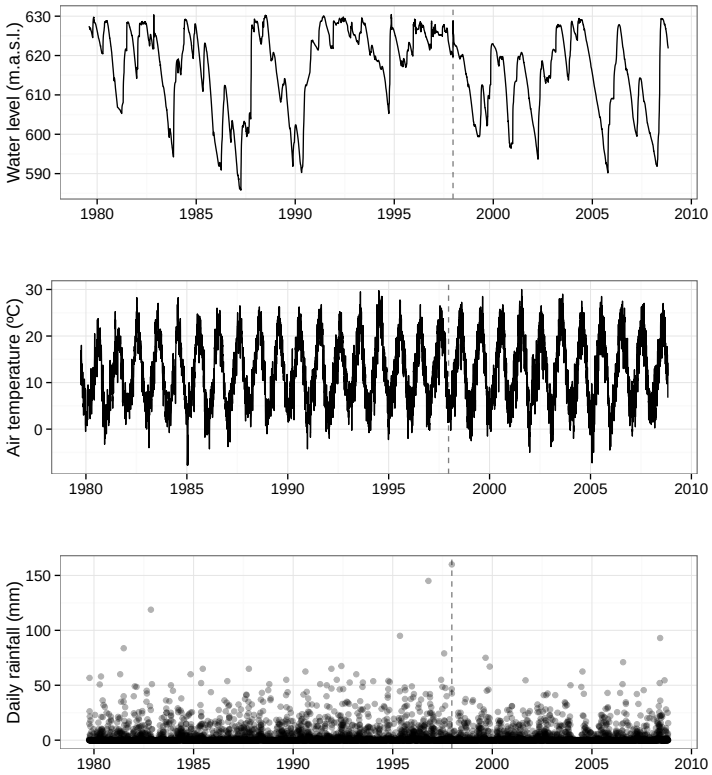

Figure 2: Time series of environmental variables at La Baells dam site. From top to bottom: water level, air temperature and daily rainfall. The vertical dashed line marks the division between training and test periods.

The predictor set includes the following 25 variables:

- Three raw environmental variables, measured at the dam site (Figure 2):

1. Air temperature
2. Reservoir level
3. Daily rainfall

- Some derived variables:

1. Average velocity of reservoir level variation in different periods prior to the measurement (10, 20 and 30 days).
2. Accumulated rainfall over various periods (30, 60, 90 and 180 days) prior to the reading.
3. Moving averages of reservoir level and air temperature over seven, 14, 30, 60, 90 and 180 days before the record.

- Time-related variables:

1. Year
2. Month
3. Number of days from the first record

The variable selection was performed according to dam engineering practice. Both displacements and leakage are strongly dependant on hydrostatic load. Air temperature is well known to affect displacements, in the form of a delayed action. It is not clear whether it strongly influences leakage or not. Whereas Seifart et *al.* [20] reported that leakage in the Itaipú Dam follows a seasonal cycle "due clearly to the thermal effect on the opening and closure of joints", other studies showed no dependency [3]. Both the air temperature and some moving averages were included in the analysis.

Hydrostatic load induces an almost immediate response of the dam, although some studies suggest that there may be also a delayed effect, specially for leakage [11], [21]. Moving averages of reservoir level were considered to capture it. The velocity of variation of reservoir level over different periods was also included, following studies that suggest the existence of an influence in dam displacements [22].

In order to account for the temporal drift of the structure, both the year and the number of days from the first record were also added.

A relatively large set of predictors was used to capture every potential effect, overlooking the high correlation among some of them. In addition, the comparison sought to be as unbiased as possible, thus all the models were built using the same inputs[4]. While it is acknowledged that this procedure may favour the techniques that better handle noisy or scarcely important variables, theoretically all learning algorithms should discard them automatically during the model fitting.

3. Methods

In this section, the algorithms chosen to build the prediction models are briefly described. Although the detailed mathematical description is beyond the scope of the paper, a short description, the most relevant features, and some key references are included. All the models were built by using the language/programming environment R [23] and some of its packages, which are cited in each section. The plots were generated with the library "ggplot2" [24].

The objective is to predict an output variable $Y \in \mathbb{R}$ based on the value of a set of predictors $X \in \mathbb{R}^p$, i.e. $Y \approx \hat{Y} = F(X)$. The observed values are denoted as $(x_i, y_i), i = 1, ..., N$, where N is the number of observations. Note that each x_i is a vector with p components, each of which is referred to as x_i^j, when necessary. Similarly, $X^j, j = 1, ..., p$ stands for each dimension of the input space.

3.1. Random forests (RF)

An RF model is a group of regression (or classification) trees [25], trained on altered versions of the training set. Given that its output is the average of the prediction of each individual tree, it is an *ensemble method*. Since RF were first

[4] with the exceptions of MARS and HST, as explained in sections 3.5 and 3.6 respectively

introduced by Breiman [26], they have become highly popular as a method to build predictive models [27]. The training process has two random components:

- Only a random subset of the input variables is considered to perform each division of the input space.
- Each tree is built using a different training set, obtained from the original data via random sampling with replacement.

The aim of adding randomness is to generate substantially different trees, so that the ensemble captures as many patterns in the training set as possible. Other interesting features of RF are the following:

- They can easily handle continuous, discrete and categorical inputs, as well as missing values.
- They can naturally model non-linear interactions.
- They avoid the need to perform cross-validation, because an unbiased estimate of the generalisation error is computed during the training process (out-of-bag (OOB) error).

Two parameters can be tuned for building an RF model: the number of covariates to consider at each split (*mtry*), and the total number of trees in the forest. Neither has significant influence on the results, according to the majority of published authors (e.g. [26], [27]).

The default value of *mtry* for regression is $p/3$, with p being the number of covariates. An RF model was fitted with the default *mtry*, and then it was increased and decreased to find the value that gives the minimum OOB error. The function *tuneRF* of the R package "randomForest" [28] was used.

All RF models entailed 500 trees, with it being checked that the OOB error is stable with that size.

3.2. Boosted regression trees (BRT)

Boosting is a general scheme to build ensemble prediction models [29]. Although various methods can be selected to form the ensemble, regression trees

are frequently chosen, and were used in this work.

The idea is to build an ensemble so that each single model (often referred to as *base learners*), is fitted to the residual of the previous ensemble. The overall prediction is calculated as the weighted sum of the outputs of the base learners (unlike RF, where the prediction of the ensemble is the average).

The algorithm includes two ingredients to avoid over-fitting:

- Each learner is trained on a random subset (without replacement) of the training set. This also decreases the computational cost.
- A regularisation (shrinkage) parameter $\nu \in (0, 1)$ is applied.

Some empirical analyses show that relatively low values of ν (below 0.1) greatly improve generalisation capability [29]. The optimal value depends on the number of base learners. In practice, it is common to set the regularisation parameter and calculate a number of trees such that the training error stabilises [30]. Subsequently, a certain number of terms are "pruned" by using for example cross-validation. The library used [30] allows choice of the number of trees by different methods. The value $\nu = 0.001$ was considered and the number of trees was selected by means of five-fold cross-validation. The process was repeated by using trees of depth 1 and 2 (*interaction.depth*), and the most accurate for each target selected.

3.3. Neural networks (NN)

NN models have been applied to solve a wide variety of problems. Among the different types of NN found in the literature [13], the multilayer perceptron (MLP) was selected for this work. An MLP is formed by a number of single units, called perceptrons, organised in different layers. The simplest architecture of an MLP was used, which involves three layers: input, hidden and output. Each perceptron in the hidden layer applies a nonlinear transformation to the inputs, and yields a response, which is later combined to compute the model output. Thus, NN are appropriate to model non-linear input-output relations.

NN stand out as one of the most popular machine learning techniques for civil engineers. Some previous applications to dam monitoring have been mentioned in section 1.2.

The package "nnet" [31] was used, which allows tuning the NN models by setting several parameters. The *size* (number of perceptrons in the hidden layer) and the *decay* (regularisation parameter) are the most influential in the results [32].

There is not a standard procedure to set the value of these parameters. Mata [12] selected the number of neurons between 3 and 30 that provided the highest accuracy via cross-validation. Hastie et *al.* [32] suggest to put down a "reasonably large" number of units and train the network with some regularisation. In this work, both criteria were combined: the search for the number of perceptrons was not so exhaustive, but their performance was assessed in combination with different degrees of regularisation. More precisely, all the possible combinations of three, 12 and 25 perceptrons (*size*) with *decay* of 0.1, 0.01 and 0.001 were tried, and the pair of values which showed the minimum error via five-fold cross-validation was chosen. For each fold, 15 NN with different initialisations were fitted, and the average error was compared. Thus, the accuracy of every combination of parameters was computed on the basis of 75 NN.

The selected parameters were applied to train 20 NN models over the whole training set with the function *avNNet*, from the R package "caret" [33]. The final prediction was computed as the average of the 20 NN.

3.4. Support vector machines (SVM)

This learning algorithm is based on a non-linear transformation of the predictor variables to a higher dimensional space (often referred to as *feature space*), and a linear regression on these transformed variables. The mathematical development of the method is complex and beyond the scope of the paper. Detailed and rigorous descriptions can be found in [34] and [35], and a recent application in predicting dam behaviour is reported in [36]. The method uses an ε-insensitive error function that neglects errors below the threshold ε. The al-

gorithm searches for a trade-off between the minimum error and the smoothness of the approximating function. As in the case of NN, the values of SVM model parameters are frequently selected via cross-validation among a subset of the possible values [17], [36], [37]. A similar criterion was followed in this work, where the library "e1071" within the R environment [38] was used to tune the most important parameters [37] of an SVM model:

- The "cost" parameter, C. Values of 10, 100 and 500 were tested.
- The width of the ε-insensitive error function, ε. The default value (0.1) was chosen.
- The kernel function, which defines the mapping from the input to the feature space. A radial basis function was considered: $K(x_i, X) = e^{-\gamma|x_i - X|^2}$
- The γ parameter of the kernel. Values of 1, 0.1, 0.01, 0.001 and 0.0001 were tried.

The 15 possible combinations of C and γ were applied to fit SVM models on the training data. Four-fold cross-validation was performed to obtain the best values. Each fold and combination of parameters was repeated five times to account for randomness, and the one with the lowest average error was selected to train an SVM model over the whole training set.

3.5. Multivariate adaptive regression splines (MARS)

MARS is an adaptive algorithm originally proposed by Friedman [39]. It is based on the combination of elementary piecewise linear functions, which definition depends on the data. As an example, an input data $x_j^j = k$ defines a pair of basic functions of the form $(X^j - k)_+$ and $(k - X^j)_+$. The subscript "+" stands for the positive part, i.e.: $(X^j - k)_+ = X^j - k$ if $X^j > k$; 0, otherwise [32]. The algorithm starts with a constant value and adds pairs of functions as long as the training error decreases above a given threshold. This is the forward pass. At the end of this step, the resulting model typically over-fits the data. Then a "pruning" process is followed, during which some of the functions are

eliminated according to the generalised cross validation (GCV) criterion. GCV is a modification of the residual sum of squares (RSS) which takes into account the number of parameters of the model [32]. In practice, the method searches for a trade-off between the reduction in the training error and the complexity of the model.

MARS models are well suited to non-linear problems, as well as easily interpretable. Furthermore, the algorithm implicitly performs variable selection, given that the functions in the final ensemble depend only on the most relevant predictors X^j.

The work was performed using the library "earth" [40], and the parameter tuned was the maximum number of terms in the final model (*nprune*). Five-fold cross-validation was run, repeated five times ($nfold = 5$ and $ncross = 5$ in the *earth* function). In principle, the model with the highest coefficient of determination (RSq) should be selected. However, the results of some preliminary tests showed that in most cases the RSq rose sharply after adding the first few terms, and remained almost constant afterwards. For the sake of model simplicity and generalisation capability, the lower value of *nprune* with $RSq \geq mean(RSq) - SD(RSq)$ was selected, a criterion similar to the *1 SE rule* proposed by Breiman et *al.* [41]. The same tests also revealed that the models with one or more time-dependant functions in the final ensemble (i.e. considering the year and/or the day since the first record) had poor generalisation ability. Therefore, both inputs were removed from the set of predictors.

3.6. HST

A conventional HST model was also built, in order to compare the results with current engineering practice. The most typical form was chosen:

$$\begin{aligned} \hat{Y} = F(t,h,s) \quad = \quad & a_0 + a_1 h + a_2 h^2 + a_3 h^3 + a_4 h^4 + a_5 h^5 \\ & + a_5 e^{-t} + a_6 t + a_7 cos(s) + a_8 sin(s) \\ & + a_9 sin^2(s) + a_{10} sin(s) cos(s) \end{aligned}$$

$$s = \frac{d}{365,25} 2\pi$$

where d is the number of days since 1 January, t is the elapsed time (years), h is the reservoir level, and $a_1, a_2, ..., a_{10}$ are the coefficients to fit.

3.7. Measures of accuracy

The accuracy of regression models is frequently measured via the mean absolute error (MAE), computed as:

$$MAE = \frac{\sum_{i=1}^{N} |y_i - F(x_i)|}{N}$$

where N is the size of the training (or test) set, y_i are the observed outputs and $F(x_i)$ the predicted values. Given that MAE is measured in the same units as the target variable, it provides a useful indication of prediction accuracy. However, it takes into account neither the mean value of the output, nor its deviation. Moreover, it is not appropriate to compare results correspondent to outputs of a different nature (e.g. displacements vs flows). To overcome these drawbacks, the study is mostly based on the average relative variance (ARV) [42]:

$$ARV = \frac{\sum_{i=1}^{N} (y_i - F(x_i))^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} = \frac{MSE}{\sigma^2}$$

where $\bar{y}$ is the output mean. Given that ARV denotes the ratio between the mean squared error (MSE) and the variance (σ^2), it accounts both for the magnitude and the deviation of the target variable. Furthermore, a model with ARV=1 is as accurate a prediction as the mean of the observed outputs.

4. Results and discussion

Models for 14 targets, with six learning algorithms, trained over four different training sets were fitted, i.e. $14 \times 6 \times 4 = 336$ models. Due to space constraints, only one plot is presented in Figure 3 as an example. It shows the predictions of the BRT model trained over the whole training set, in comparison with the measured data for three targets of different kind (P1DR1, P1DT1

and AFMD50PR). It provides an intuition on the goodness of fit achieved, and highlights how the ARV allows comparison of the accuracy between different targets. Although the highest MAE corresponds to P1DR1, it yields the lowest ARV at the same time, because of its higher variance.

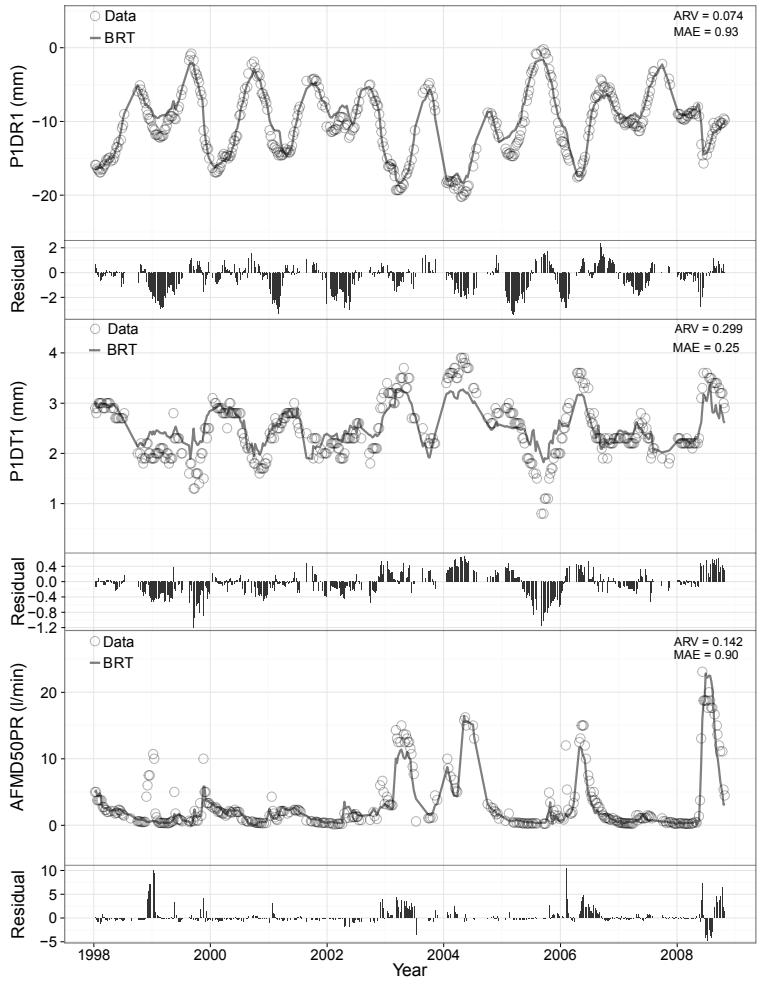

Figure 3: Measured data (circles) versus predictions of the BRT model (lines) for the test period. The residuals between them are included below each plot. From top to bottom: P1DR1, P1DT1 and AFMD50PR.

It is commonly accepted that increasing the amount of training data leads to

Type	Target	RF	BRT	NN	SVM	MARS	HST
Radial (mm)	P1DR1	1.70	0.93	**0.58**	0.75	2.32	1.35
	P1DR4	1.05	**0.71**	**0.68**	0.76	1.50	1.37
	P2IR4	0.94	0.97	1.02	1.05	**0.85**	1.12
	P5DR1	0.86	**0.70**	**0.64**	1.35	0.89	0.88
	P6IR1	1.47	0.69	0.72	**0.60**	1.67	0.91
Tangential (mm)	P1DT1	**0.24**	**0.25**	0.52	0.35	0.55	0.47
	P1DT4	**0.15**	**0.15**	0.18	0.19	0.22	0.20
	P2IT4	0.13	**0.11**	0.13	0.12	0.14	**0.10**
	P5DT1	0.40	0.22	**0.19**	0.38	0.47	**0.18**
	P6IT1	**0.28**	**0.27**	0.39	0.94	0.39	0.51
Leakage (l/min)	AFMD50PR	1.24	**0.90**	2.11	4.25	1.74	2.24
	AFMI90PR	0.18	0.15	**0.07**	0.33	0.25	0.28
	AFTOTMD	1.82	**1.60**	3.04	5.38	1.85	2.60
	AFTOTMI	0.91	**0.42**	0.83	1.49	1.49	1.11

Table 2: MAE for each output and model, fitted on the whole training set (18 years). The values within 10% from the minimum are highlighted in bold, and the minimum MAE are also underlined. The results correspond to the test set.

a better model performance. However, this may not be the case of dams, where certain drift is often observed. A first set of calculations was done with all the available data (18 years) for comparison purposes. The results are presented in Table 2, which contains the MAE for each target and model. The effect of the training set size was also examined, as described later in this section.

It can be seen that models based on ML techniques mostly outperform the reference HST method. NN models yield the highest accuracy for radial displacements, whereas BRT models are better on average both for tangential displacements and leakage flow. It should be noted that the MAE for some tangential displacements is close to the measurement error of the device ($\pm 0.1mm$).

Figure 4 shows the results in terms of ARV for each model and type of

output. It should be remembered that models with ARV > 1 can be considered as being of little use. The error is lower for radial displacements, whereas there is not a great difference between the ARV for leakage flow and tangential displacements. These results are in accordance with engineering knowledge: the prediction of tangential displacements is more difficult because the signal-to-noise ratio is lower than for radial displacements (while the measurement error is the same, the standard deviations are highly different, as shown in Table 1). The measurement error for leakage flow is negligible, but it is governed by a more complex physical process, which makes it harder to predict.

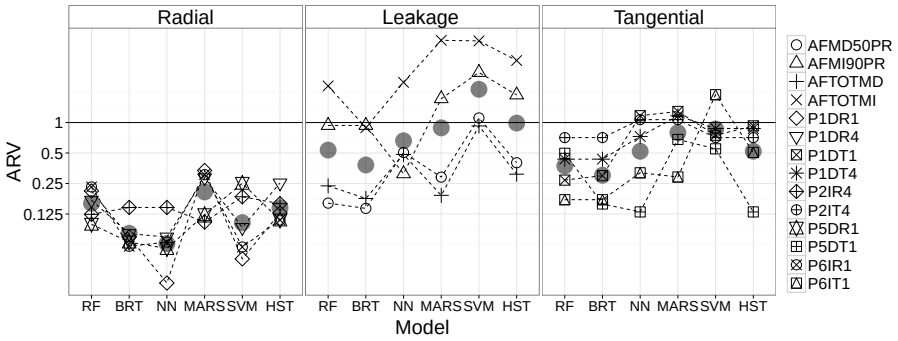

Figure 4: *ARV* for each target and model, fitted on the whole training set (18 years). Models with *ARV* > 1.0 are less accurate than the sample mean. The average values for each technique and type of output are plotted with black dots. Note the logarithmic scale of the vertical axis. The results correspond to the test set.

The study was repeated with each technique, using different training set sizes, namely five, 10 and 15 years. The results were compared to those obtained previously, with 18 years. The test set was the same as before (1998-2008). Figure 5 shows the results. An important decrease in error is observed in most cases between models trained on five and 10 years. This decrease is dramatic for HST (note that some of the results for HST and five years lie outside the vertical limit of the plots).

Although some previous studies offered similar results [19], in this case such effect may be more pronounced due to the fact that the reservoir level remained

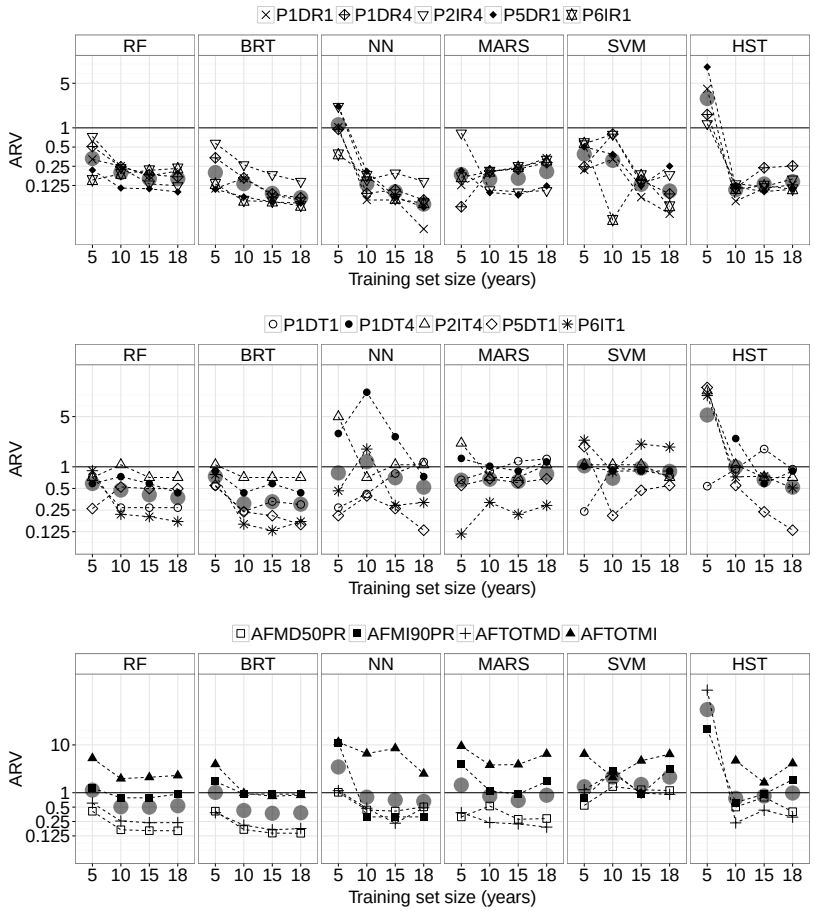

Figure 5: ARV for each model and training set size. Top: radial displacements. Middle: tangential displacements. Bottom: leakage flow. Some HST models trained over 5 years are out of the range of the vertical axis, thus highly inaccurate. The results correspond to the test set.

high in the 1993-1998 period (Fig. 2). Models fitted on those years have no information on the dam behaviour when the reservoir is at low levels, and therefore the prediction of the dam response in such situations may be highly inaccurate.

When increasing the training set up to 15 and 18 years, the variation is either negligible (i.e. BRT models for leakage, Figure 5, bottom), or there is a small decrease in error (i.e. NN models for radial displacements, Figure 5,

top). In some cases, the error even increases, such as in HST models for radial displacements (Figure 5, top). Some techniques do not show a clear trend, such as MARS models for tangential displacements (Figure 5, bottom).

Table 3 compares the best models overall with those trained on the entire training set (18 years). Although the use of the whole training set is optimal

Target	Best model 18 years	MAE 18 years	Best model overall	MAE overall	Best training size (years)	MAE reduction (%)
P1DR1	NN	0.58	-	-	-	-
P1DR4	NN	0.68	MARS	0.60	5	13.3
P2IR4	MARS	0.85	MARS	0.81	15	4.7
P5DR1	NN	0.64	-	-	-	-
P6IR1	SVM	0.60	SVM	0.53	10	11.7
P1DT1	RF	0.24	BRT	0.22	10	8.3
P1DT4	RF/BRT	0.15	BRT	0.14	10	6.7
P2IT4	HST	0.10	-	-	-	-
P5DT1	HST	0.18	-	-	-	-
P6IT1	BRT	0.27	MARS	0.23	5	14.8
AFMD50PR	BRT	0.90	BRT	0.89	15	1.1
AFMI90PR	NN	0.07	-	-	-	-
AFTOTMD	BRT	1.60	BRT	1.57	15	1.9
AFTOTMI	BRT	0.42	-	-	-	-

Table 3: Comparison between the best models fitted using the whole training set and the best overall. Empty rows correspond to outputs for which no improvement is achieved by using a smaller training set. The results correspond to the test set.

for six out of 14 targets, significant improvements are reported in some cases by eliminating some of the early years. Surprisingly, for two of the outputs, the lower MAE corresponds to a model trained over five years, which in principle was assumed to be too small a training set. MARS is especially sensitive to the size of the training data. The MARS models trained on five years improve the accuracy for P1DR4 and P6IT1 by 13.3 % and 14.8 % respectively.

These results strongly suggest that it is advisable to select carefully the most appropriate training set size. This should be done by leaving an independent

validation set.

5. Summary and conclusions

It was found that the accuracy of currently applied methods for predicting dam behaviour can be substantially improved by using ML techniques.

The sensitivity analysis to the training set size shows that removing the early years of dam life cycle can be beneficial. In this work, it has resulted in a decrease in MAE in some cases (up to 14.8%). Hence, the size of the training set should be considered as an extra parameter to be optimised during training.

Some of the techniques analysed (MARS, SVM, NN) are more susceptible to further tuning than others (RF, BRT), given that they have more hyper-parameters. As a consequence, the former might have a larger margin for improvement than the latter.

A more careful selection of variables could also improve the fit. It should be noted, though, that variable selection is an issue in itself, and will be the subject of further work. It may not only decrease the error, but also help to build more understandable models.

However, both detailed tuning and careful variable selection increase the computational cost and complicate the analysis. If the objective is the extension of these techniques for the prediction of a large number of variables of many dams, the simplicity of implementation is an aspect to be considered in model selection.

In this sense, BRT showed to be the best choice: it was the most accurate for five of the 14 targets; easy to implement; robust with respect to the training set size; able to consider any kind of input (numeric, categorical or discrete), and not sensitive to noisy and low relevant predictors.

However, none of the algorithms provided the highest accuracy in all cases. Therefore, if the main objective is to achieve the best possible fit, the analysis should not be limited to a single technique.

It seems clear that the models based on ML algorithms are more suitable to

reproduce non-linear effects and complex interactions between input variables and dam response.

Nonetheless, these tools must be employed rigorously, given their relatively high number of parameters and flexibility, what makes them susceptible to overfit the training data. It is essential to check their generalisation capability on an adequate validation data set, not used for fitting the model parameters.

Regardless of the technique used, engineering judgement based on experience is critical for building the model, for interpreting the results, and for decision making with regard to dam safety.

6. Acknowledgements

The authors thank the Catalan Water Agency, owner of La Baells dam, and the company Ofiteco for providing the monitoring data.

The research has been partially supported by the Spanish Ministry of Economy and Competitiveness (*Ministerio de Economía y Competitividad*, MINECO) through the projects iComplex (IPT-2012-0813-390000) and AIDA (BIA2013-49018-C2-1-R and BIA2013- 49018-C2-2-R).

References

[1] Swiss Committee on Dams, Methods of analysis for the prediction and the verification of dam behaviour, Tech. rep., ICOLD (2003).

[2] L. Chouinard, V. Roy, Performance of statistical models for dam monitoring data, in: Joint International Conference on Computing and Decision Making in Civil and Building Engineering, Montreal, June, 2006, pp. 199–207.

[3] A. Simon, M. Royer, F. Mauris, J. Fabre, Analysis and interpretation of dam measurements using artificial neural networks, in: 9th ICOLD European Club Symposium, Venice, Italy, 2013.

[4] G. Lombardi, F. Amberg, G. Darbre, Algorithm for he prediction of functional delays in the behaviour of concrete dams, Hydropower and Dams (3) (2008) 111–116.

[5] International Commission on Large Dams, Dam surveillance guide, Tech. Rep. B-158, ICOLD (2012).

[6] F. Restelli, Systemic evaluation of the response of large dams instrumentation. Application at El Chocón Dam, in: 9th ICOLD European Club Symposium, Venice, Italy, 2013.

[7] M. Tatin, M. Briffaut, F. Dufour, A. Simon, J. Fabre, Thermal displacements of concrete dams: Finite element and statistical modelling, in: 9th ICOLD European Club Symposium, Venice, Italy, 2013.

[8] S. Bonelli, K. Radzicki, Impulse response function analysis of pore pressure in earthdams, European Journal of Environmental and Civil Engineering 12 (3) (2008) 243–262.

[9] F. Li, Z. Wang, G. Liu, Towards an error correction model for dam monitoring data analysis based on cointegration theory, Structural Safety 43 (2013) 1220. doi:10.1016/j.strusafe.2013.02.005.

[10] F. Riquelme, J. Fraile, D. Santillan, R. Moran, M. Toledo, Application of artificial neural network models to determine movements in an arch dam, in: 2nd International Congress on Dam Maintenance and Rehabilitation, Zaragoza, Spain, 2011, pp. 117–123.

[11] D. Santillán, J. Fraile-Ardanuy, M. Toledo, Seepage prediction in arch dams by means of artificial neural networks, Water Technology and Science V (3), accepted for publication. [In Spanish].

[12] J. Mata, Interpretation of concrete dam behaviour with artificial neural network and multiple linear regression models, Engineering Structures 3 (3) (2011) 903 – 910. doi:10.1016/j.engstruct.2010.12.011.

[13] C. M. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995.

[14] V. Ranković, N. Grujović, D. Divac, N. Milivojević, A. Novaković, Modelling of dam behaviour based on neuro-fuzzy identification, Engineering Structures 35 (2012) 107–113. doi:10.1016/j.engstruct.2011.11.011.

[15] L. Piroddi, W. Spinelli, Long-range nonlinear prediction: a case study, in: 42nd IEEE Conference on Decision and Control, Vol. 4, IEEE, 2003, pp. 3984–3989.

[16] V. Saouma, E. Hansen, B. Rajagopalan, Statistical and 3d nonlinear finite element analysis of Schlegeis dam, Tech. rep., University of Colorado (2001).

[17] M. Herrera, L. Torgo, J. Izquierdo, R. Pérez-García, Predictive models for forecasting hourly urban water demand, Journal of Hydrology 387 (1) (2010) 141–150.

[18] G. Lombardi, Advanced data interpretation for diagnosis of concrete dams, Tech. rep., CISM (2004).

[19] A. De Sortis, P. Paoliani, Statistical analysis and structural identification in concrete dam monitoring, Engineering Structures 29 (1) (2007) 110–120.

[20] L. Seifard, A. Szpilman, C. Piasentin, Itaipú structures. Evaluation of their performance, in: 15th ICOLD Congress, 1985, pp. 287–317, Q56-R15.

[21] Q. Guedes, P. Coelho, Statistical behaviour model of dams, in: 15th ICOLD Congres, 1985, pp. 319–334, Q56-R16.

[22] F. J. Sánchez Caro, Dam safety: contributions to the deformation analysis and monitoring as an element of prevention of pathologies of geotechnical origin, Ph.D. thesis, UPM, [In Spanish] (2007).

[23] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2013).
URL http://www.R-project.org/

[24] H. Wickham, ggplot2: elegant graphics for data analysis, Springer New York, 2009.

[25] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and regression trees, Wadsworth & Brooks, Monterrey, CA, 1984.

[26] L. Breiman, Random forests, Machine learning 45 (1) (2001) 05–32.

[27] R. Genuer, J. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recognition Letters 31 (14) (2010) 225 – 236. doi:10.1016/j.patrec.2010.03.014.

[28] A. Liaw, M. Wiener, Classification and regression by randomForest, R news 2 (3) (2002) 18–22.

[29] J. Friedman, Greedy function approximation: a gradient boosting machine, Annals of Statistics (2001) 1189 – 1232.

[30] G. Ridgeway, Generalized Boosted Models: A guide to the gbm package, R package vignette (2007).
URL http://CRAN.R-project.org/package=gbm

[31] W. N. Venables, B. D. Ripley, Modern Applied Statistics with S, 4th Edition, Springer, New York, 2002, iSBN 0-387-95457-0.

[32] T. Hastie, R. Tibshirani, J. Firedman, The Elements of Statistical Learning - Data Mining, Inference, and Prediction, 2nd Edition, Springer, New York, 2009.

[33] M. Kuhn, Building predictive models in R using the caret package, Journal of Statistical Software 28 (5) (2008) 1–26.
URL http://www.jstatsoft.org/v28/i05/paper

[34] A. J. Smola, B. Schlkopf, A tutorial on support vector regression, Statistics and computing 14 (3) (2004) 199–222.

[35] J. M. Moguerza, A. Muñoz, Support vector machines with applications, Statistical Science (2006) 322–336.

[36] V. Ranković, N. Grujović, D. Divac, N. Milivojević, Development of support vector regression identification model for prediction of dam structural behaviour, Structural Safety 48 (2014) 33–39.

[37] B.-J. Chen, M.-W. Chang, et al., Load forecasting using support vector machines: A study on eunite competition 2001, IEEE Transactions on Power Systems 19 (4) (2004) 1821–1830.

[38] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, r package version 1.6-1 (2012).

[39] J. H. Friedman, Multivariate adaptive regression splines, The Annals of Statistics (1991) 1–67.

[40] S. Milborrow., earth: Multivariate Adaptive Regression Spline Models, R package version 3.2-6 (2013).
URL `http://CRAN.R-project.org/package=earth`

[41] L. Breiman, J. Friedman, R. Olshen, C. Stone, D. Steinberg, P. Colla, Classification and Regression Trees, Wadsworth: Belmont, CA, 1984.

[42] A. S. Weigend, B. A. Huberman, D. E. Rumelhart, Predicting sunspots and exchange rates with connectionist networks, in: S. Eubank, M. Casdagli (Eds.), Proc. of the 1990 NATO Workshop on Nonlinear Modeling and Forecasting (Santa Fe, NM), Vol. 12, Addison-Wesley, Redwood, CA, 1992, pp. 395–432.