Prediction of Dam Behavior Based on Machine Learning Methods

Patricia Alocén & Miguel Á. Fernández-Centeno
ACIS Innovation+Engineering S.L. (ACIS2in), Parla, Madrid, Spain, 28983.

Universidad Politécnica de Madrid (UPM), ETS de Ingenieros de Caminos, Canales y Puertos; Madrid, Spain, 28040.

Miguel Á. Toledo
Universidad Politécnica de Madrid (UPM), ETS de Ingenieros de Caminos, Canales y Puertos; Madrid, Spain, 28040.

ABSTRACT: Prediction of dam behavior plays an important role in the field of dam safety as it can be used to establish warning levels and detect dam failure. Recently, Machine Learning techniques have been increasingly applied in this field due to their success in other areas. Our methodology is based on such techniques to predict different measurements of and arch dam's behavior and analyze the influence of external conditions. First, we performed an exploratory analysis and selected the most important variables for prediction. We measured the degree of similarity between external factors in the available years. Then, several models were trained for each target variable, and the optimal was selected, which were used to make short- and long-term predictions and determine warning levels. The results show that the short- and long-term moving averages of the water level are the most important variables regarding the prediction of the displacement and different groups of years in external conditions were also observed. SVM, NN, and BRT were the most accurate methods and their errors were used to determine the warning levels.

1 INTRODUCTION

The practical problem to be solved in this research is denoted as Theme_A and its main objective is to predict the behavior of a double curvature arch dam. In our case, we chose to use Machine Learning techniques to configure models and perform the relevant analysis.

The developers of Acis2in applied algorithms developed in other research projects to solve this practical problem, some of them implemented in our web application called SmartDam. We analyzed the external conditions affecting the dam and trained the Machine Learning models to make the short- and long-term predictions required.

2 METHODOLOGY

2.1 *Exploratory analysis*

The research carried out in this workshop began with an exploratory analysis of the external factors and target variables. Their series and distributions were studied to understand their individual behavior, as well as the relationships between them.

To perform the individual analysis of the variables, we have plotted their time series, density, and boxplots grouped by the available years in the data set. The latter two were used to observe their mean, dispersion, and range. We observed the difference between the behavior of the variables in each of the years through these graphs. The emphasis was placed on the study of water level, which, as discussed in the results section, has a crucial role in the prediction of the target variables.

Correlation and scatterplot graphs were used to analyze the relationship between target variables and external factors. The former shows values between 0 and 1 indicating the degree of linear relationship between them, while the latter shows the type of relationship they hold (linear, non-linear, etc.).

2.2 *Synthetic variables*

The next step in our methodology was to calculate synthetic variables related to the past of external factors. These variables play an important role in the training process, since the effect of external factors does not immediately affect the dam, but rather there is a delayed effect.

Three types of variables of different orders were calculated: moving averages (MM), aggregates (AG), and variation ratio (VEL).

Assuming we have a time series of variable $X \in (1 \text{ x } m)$, where t is the instance at time t of variable X, the synthetic variables are computed as follows:

$$\text{X\_MM}_{t,k} = \frac{1}{k} \sum_{i=1}^{k} x_{t-i}\,; \tag{1}$$

$$\text{X\_AG}_{t,k} = \sum_{i=1}^{k} x_{t-i}\,; \tag{2}$$

$$\text{X\_VEL}_{t,k} = \frac{x_t - x_{t-k}}{k}; \tag{3}$$

where k is the order of the synthetic variable and t is the variable X at time t.

Short- and long-term synthetic variables of orders 7, 15, 30, 30, 60, 90, 180, and 365 were calculated. Among these variables, those of greatest importance in the prediction of the target variables were selected using our variable selection algorithm explained below.

2.3 *Variable Selection*

Variable selection arises due to the need to reduce the dimensions of the large data set generated after calculating the synthetic variables. Logically, all these variables are closely related to each other. Therefore, it is important to select only those that provide relevant information to the model to improve accuracy and reduce computational cost.

The outline of the selection algorithm is as follows:

1. Calculation of the degree of importance through Support Vector Machine (SVM).
2. Sort variables by importance degree in descending order.
3. Selection of certain numbers of variables: 10%, 20%, 30%, 50%, 60% and 80%.
4. Execute an SVM for each quantity in Step 3.
5. Selection of the variables that generate the most accurate model.

First, a simple model was trained using SVM to calculate the degree of importance of each variable by measuring the area under the ROC curve. In our experience in other research, SVM takes less time to run and often gives the same results as other variable selection methods, such as ensembles of decision trees.

The next question to be answered was how many variables should be used to optimize the accuracy of the final model. The most accurate selection methods, such as leave-one-out, may become computationally expensive if the dimensions of the training set are large. Therefore, in our algorithm, different percentages of variables are selected in descending order of importance, and a simple model is trained with SVM for each of these quantities. Finally, the quantity that gives the smallest error is selected.

The variables resulting from this last step were used to train the final model to predict the target variable, where a search for the optimal hyperparameters and an estimation of the error was performed through cross-validation.

2.4 *Similarity between the external conditions of the years*

To further study the external factors that influence the dam, an analysis of the similarity of these across the available years has been developed. The similarity measure used is summarized by calculating the Euclidean distance of the instances to the centroids of the Principal Components of the training years.

The variables used as inputs in the algorithm are the important variables resulting from the previous section. Therefore, because each target variable has a different set of important variables, the results of this section may differ among the target variables, even though the original external conditions are the same for all of them. It is important to keep this in mind when analyzing the results written in the next chapter.

The steps taken to calculate the similarity between the external conditions of each year are detailed below:

1. Data scaling: All variables are converted to the same scale.
2. Calculation of Principal Components (PC).
3. The *n* PCs that explain 90% of the variability of the set were selected. Hence, we have an *m* x *n* data matrix, where *m* is the number of instances (or available dates) and *n* the number of principal components selected.

$$PCM = \begin{pmatrix} pc_{11} & \cdots & pc_{1n} \\ \vdots & \ddots & \vdots \\ pc_{m1} & \cdots & pc_{mn} \end{pmatrix}; \quad (4)$$

4. The centroids of each PC were calculated for each year. Then, we have a vector of coordinates C in the principal component space for each year available in the data set.

$$C_y = \left(\overline{PCM}_{m\in y,1}, \overline{PCM}_{m\in y,2}, \ldots, \overline{PCM}_{m\in y,n}\right); \quad (5)$$

where $\overline{PCM}_{m\in y,1}$ is the mean of the first principal component of the instances belonging to the year y and $y \in [1, \ldots, k]$, being k the total number of available years.

$$\overline{PCM}_{m\in y,j} = \frac{1}{m_y}\sum_{i=1}^{m_y} PCM_{i,j}; \quad (6)$$

where j is the principal component taking values from 1 to n and m_y is the number of instances in year y.

5. Clusters of centroids were identified using the kmeans algorithm. Consequently, we obtained the groups of the most similar centroids.
6. The Euclidean distance from each point of the *PCM* matrix to each centroid of the training years ($y \in training$) was calculated, excluding the year to which the instance belongs ($i \notin y$).
    a. Euclidean distance from each point of the *PCM* matrix to the centroid of the year y:

$$d_{E_i}(PCM, C_y) = \sqrt{\sum_{j=1;i\notin y}^{n} \left(PCM_{i,j} - \overline{PCM}_{m\in y,j}\right)^2}; \quad (7)$$

where $i \notin y$ and j is the Principal Component ($j \in [1, \dots, n]$).

Consequently, we have a vector 1 x k (years) for each i instance. That makes up the following matrix:

$$D_E = \begin{pmatrix} d_{E_{11}} & \cdots & d_{E_{1k}} \\ \vdots & \ddots & \vdots \\ d_{E_{m1}} & \cdots & d_{E_{mk}} \end{pmatrix}; \quad (8)$$

Finally, the variable measuring distance was computed as the average of the distances to all centroids.

$$\overline{d_{E_i}}(PCM, C) = \frac{1}{k}\sum_{y=1}^{k} d_{E_i}(PCM, C_y) = \frac{1}{k}\sum_{y=1}^{k} D_{E_{iy}}; \quad (9)$$

Thus, we have a variable $\overline{d_E}(PCM, C)$ of dimension 1 x m that measures the average distance from each point to the centroids of the training years.

7. The average per year of the variable generated in the previous step was calculated. This variable is the average distance from each year to the rest of them (belonging to training).

$$dE_y = \frac{1}{m_y}\sum_{i=1,i\in y}^{m_y} \overline{d_{E_i}}(PCM, C); \quad (10)$$

where m_y is the number of years of y.

The result is a vector of dimensions *1 x k*: $dE = (dE_1, \dots, dE_k)$, which indicates how far is each year to the resto of the training years.

8. The minimum and maximum distance of the points of each year to the centroids of the training years were calculated as the minimum and maximum distance of each year to the centroids of the rest:

$$\min_{\substack{y,k\in(1,\dots,k)\\ k\neq y}} d_E(PCM_y, C_k); \quad \max_{\substack{y,k\in(1,\dots,k)\\ k\neq y}} d_E(PCM_y, C_k); \quad (11)$$

where $d_E(PCM_y, C_k)$ is the average distance of the *PCM* instances belonging to year y, and the centroid of year k. This step indicates the nearest and furthest training year in external conditions for each of the years.

It should be noted that the distances from the points to the centroids are made considering only those of the years used during training. If the instance belongs to one of the training years, the centroid of that year is excluded from the calculation of the distances. For example, since 2003 was used to train the model, the distances of the 2003 instances were calculated by measuring the Euclidean distances to the centroids of the training years except 2003. However, since 2016 was not used for training, its distances were calculated considering all centroids of the training years.

2.5 *Training and evaluation of models*

The model training stage consisted of the selection and training of models of different nature. Methods that are potentially accurate based on previous research experience were selected:

- Boosted Regression Trees (BRT).
- Random Forest (RF).
- Support Vector Machine (SVM).
- Neural Network (NN).
- Generalized Linear Regression (LM).
- Bayesian Neural Network (RRBB).
- Hydrostatic-Season-Time (HST).

Cross-validation was used to evaluate the models and estimate the optimal hyperparameters for each case. In this research, the folds match the years available in the dataset, which correspond to the dam cycles. The estimated error by averaging the error across folds ($RMSE_{cv}$) is more robust than the RMSE over the validation set (year 2012) because it uses more dates.

Therefore, the error measures used in this methodology are the *RMSE* of the CV ($RMSE_{cv}$) and the RMSE of validation ($RMSE_{val}$). The mathematical form of the *RMSE* is as follows:

$$RMSE = \sqrt{\sum_{i=1}^{m} \frac{(\hat{y}_i - y_i)^2}{m}}; \tag{12}$$

where m is the total number of records in the data set, $\hat{y}$ the predicted values and y the actual values.

Considering that k years are available, we have an *RMSE* for each k years:

$$RMSE_{cv} = \frac{1}{k}\sum_{j=1}^{k} RMSE_j\,; \tag{13}$$

The measure $RMSE_{val}$ is simply the *RMSE* over the validation year, 2012 in this case.

The optimal hyperparameters of each model were selected by searching the combination that gives the lowest error on average. For each combination, a model was created for each fold; then the average *RMSE* committed across the folds was calculated and the combination with the lowest error was selected.

Accordingly, we obtained an estimated error for each of the seven trained models. The last step of this stage was to select the optimal model, which was the one with the lowest value of $RMSE_{cv}$.

2.6 *Warning levels*

Once the optimal models were selected for each target variable, the warning levels were generated. These limits were determined on the estimated error of the prediction model for each component and application segment.

It is believed necessary to add to our prediction a component of the error committed by the model, since there is always a margin of error in the prediction. One of the possibilities considered was to add to the prediction value the average error of the model. However, it was decided to establish different confidence or sensitivity coefficients that multiply the standard deviations of the error in each case. Hence, different confidence bands are generated along the prediction span.

The formula for the warning levels is as follows:

$$U_{\pm} = \hat{y} \pm c\sigma_e; \tag{14}$$

where $\hat{y}$ is the predicted value, c is the chosen coefficient and σ_e the standard deviation of the error.

Figure 1. Example of error density with the mean and coefficient points multiplied by the error variance.

The coefficient selected to establish the warning level was 2, since this interval approximately holds 95% of the real values.

3 RESULTS AND DISCUSSION

This chapter presents and discusses the results obtained by applying the methods previously presented. The subdivisions of the Methodology section are similarly implemented in this chapter for ease of understanding and to present the results in an orderly way.

3.1 *Exploratory analysis*

An exploratory analysis of external factors was carried out to determine their behavior and relationship with the target variables.

Figure 2 shows a cyclical behavior of the water level series. What is striking here is the pronounced water level drops observed in 2003, 2006 and 2016. After the decrease in water level in 2006, its average in the following years is higher due to higher minimum values. They progressively decrease in average until the drop of 2016, which makes it an unusual year compared to the past.

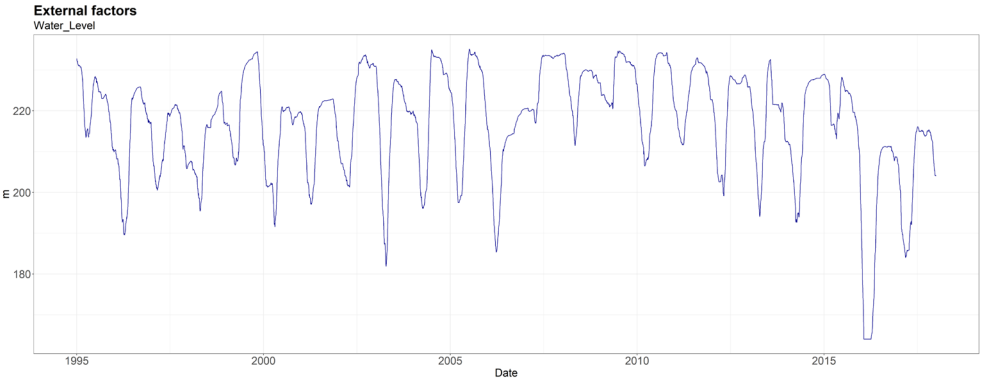

Figure 2. Water Level series over time

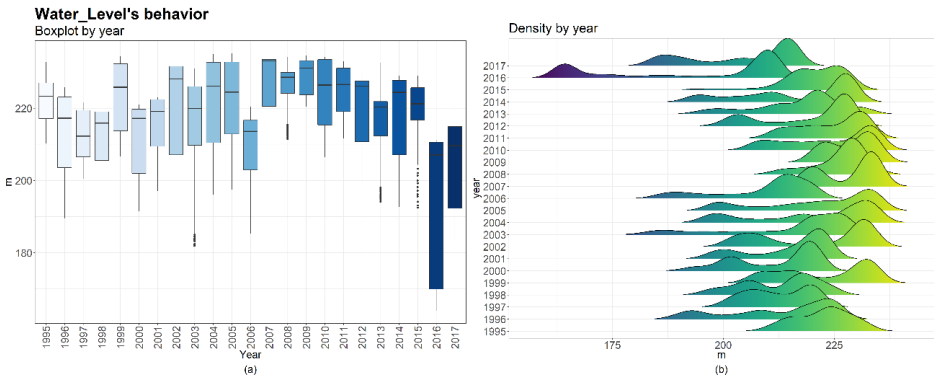

Figure 3. Boxplots (a) and density plots (b) of Water Level by years.

The plot (a) in Figure 3 shows what appears to be different behaviors according to the amplitude and the median value of the boxes: from 1995 to 2005, where the minimum values of water level are low; from 2007 to 2015, where values are higher; and the atypical periods such as 2006, 2016 and 2017. The amplitudes of years from the first period mentioned are similar, although the values of the water level vary, especially those belonging to 2003, where a larger decrease is observed. This event makes the lower whisker longer and, accordingly, outliers appear. Year 2006 is significantly different from the immediately preceding years, as the amplitude of its box is smaller, implying that the water level lays within a narrower range. From that year on, there is an increase in the water level, where we find higher medians and values that progressively decrease. Undoubtedly, the most atypical period is 2016, where the lowest values are found.

These remarks are also seen in the density graph (Figure 3, b), where several averages are observed due to cyclical rises and drops in the water level. Most of the years have similar means, except 1996, 2003,2006, 2016 and 2017, which have lower mean of minimum values than the others. The most unique year also in this type of graph is 2016, where a particularly steeper water level drawdown is observed.

This behavior contrasts with the scarce temporal variation of temperatures, whose series show the usual cyclical behavior, and very similar means and medians over the years were observed.

Regarding the pendulum series, some changes can be identified, which might be related to the previously mentioned water level drops.

The variable most correlated with both pendulums is the water level, with values 0.62 (CB_236_196) and 0.9 (CB3_195_161). Temperature, on the other hand, has a smaller linear relationship with both pendulums, finding its maximum at |-0.37| (CB_236_196). This coincides with the results of the analysis made in the previous paragraphs, where the variation of the pendulum values seemed to be related to the water level.

3.2 *Synthetic variables*

Once the exploratory analysis was performed, the synthetic variables of the external factors were calculated to be used as inputs in the modeling training.

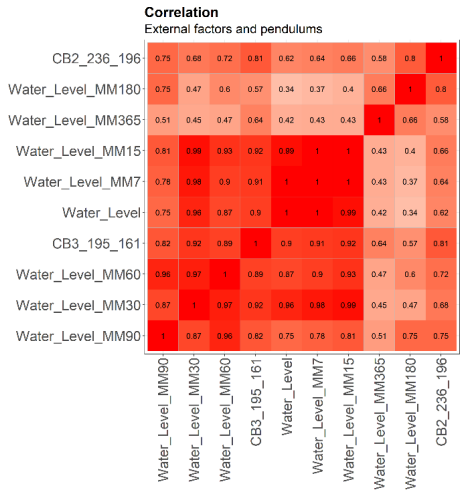

Figure 4. Correlation plot of Water Level moving averages and pendulums.

The correlation plot (Figure 4) shows that some of the moving averages of the water level are more correlated with both pendulums than the original variable. This increase in correlation helped achieve accurate models.

Short-term moving averages (MM15, MM30, etc.) are most correlated with CB3_195_161, whereas long-term averages (MM180, MM90, etc.) are most correlated with CB_236_196.

The following images show the series of these variables and their relationship to the pendulums (Figure 5 and Figure 6).

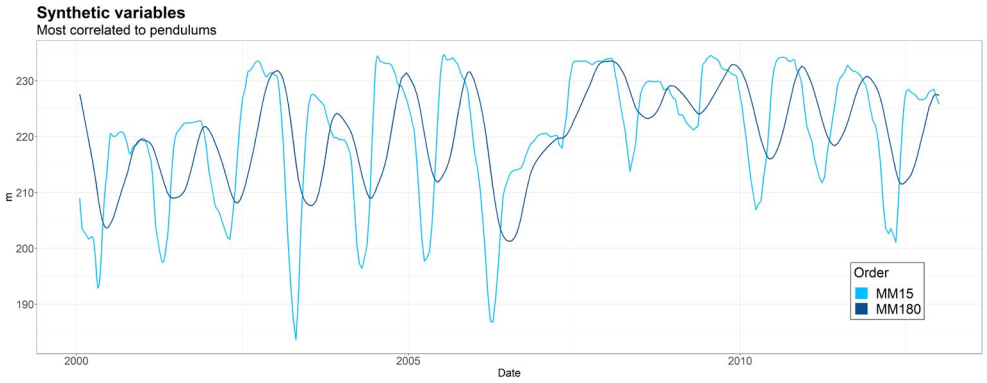

Figure 5. Series of Water Level 15 and 180 order moving averages.

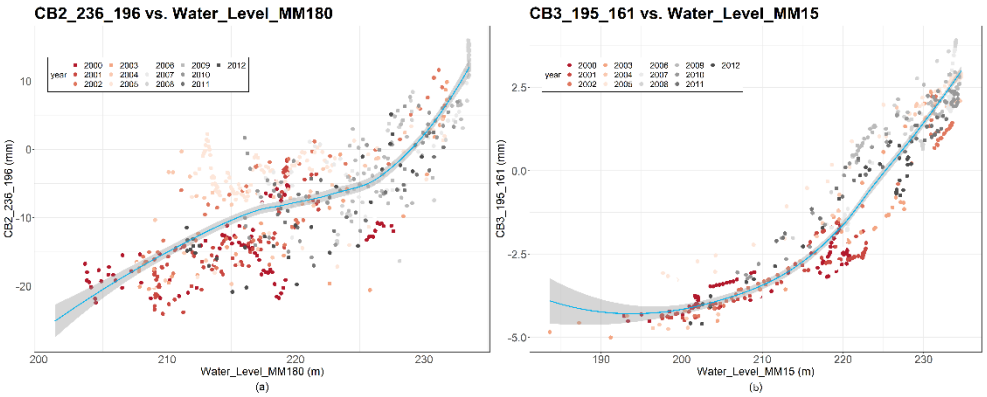

Figure 6. Scatterplot of the relationship between pendulums and moving averages 15 and 180 of the reservoir level by years: (a) CB_236_196 and (b) CB3_195_161.

Figure 6 shows a high degree of linear relationship between dam displacement of the dam and water level moving averages. Generally, points corresponding to the most current years, in gray, are concentrated in the upper right part of the graph, where the values of water level and displacement are higher. Those belonging to 2000 and 2001, in dark red, have lower values, whereas the rest are more dispersed. Given the greater dispersion in the upper pendulum compared to the water level, it would appear that it has a greater dependence on other variables than the lower pendulum in which this dispersion is smaller.

3.3 *Most important variables*

As mentioned in the Methodology section, the selection of the most important variables for each pendulum is important to increase predictive power and reduce the dimensions of the data set.

Logically, variables that have a high linear relationship will be important for the prediction of the target variable because many models tend to prefer this type of relationship for ease of modeling. This is the case with our variable selection algorithm that employs an SVM for the calculation of the importance degree.

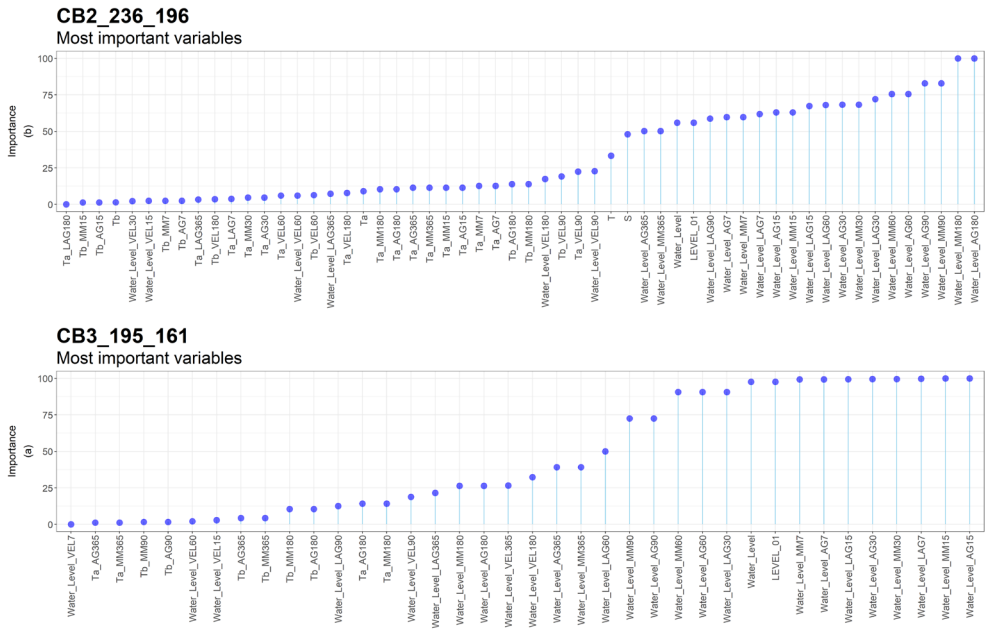

Figure 7. Most important variables of the model and their degree of importance (%).

The variables that top the list of importance for the CB_236_196 pendulum are the long-term synthetic variables (Figure 7). In contrast, in the case of CB3_195_161 the water-level short-term variables occupy this position. Temperatures are not as important, and time (T) and seasonality (S) are only ranked as important variables in the first pendulum.

It seems rare that the displacement due to water level depends more on longer term synthetic variables. This could be related to the fact that the temperature-related part of the displacement is the main component in the physical sense, not in the model. Since most of the response is delayed between 90 and 180 days, the water level information integrated at that time frame makes the extreme value correspondences match the response better.

3.4 *Similarity between years*

The possible existence of groups depending on the water level emerged in the explanatory analysis chapter. In this section, we go into detail on this issue, trying to group the different years according to their set of external factors. If these groups could be identified, the accuracy could be improved if the models were clustered by them.

As explained in the Methodology chapter, a dimension reduction of the most important

variables of both pendulums was performed by Principal Components Analysis. This is important since one may wonder why the values or groupings change from one pendulum to the other. The answer is that the calculations are made with different variables in each case (Figure 7).

The clusters resulting from running the kmeans algorithm (Figure 8) seem to coincide with the clusters that could apparently be formed by looking at the water level graphs (Figure 3). The rarest external conditions are found in 2016 and 2017, which form cluster 3.

The same groups are found for both pendulums, except for year 2002. It should be noted that the groups were made considering the centroids of 5 principal components, but to facilitate the explanation, they are represented in 2 dimensions. Hence, the actual cluster may not match what appears to be according to the graph (Figure 8).

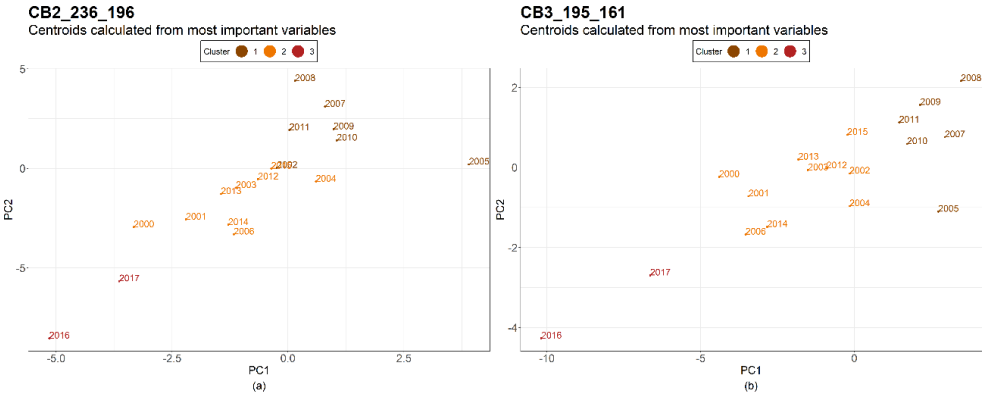

Figure 8. Centroids of the Principal Components of the years available in the dataset grouped by clusters generated through the kmeans algorithm.

Figure 8 only provides an idea of the years that are most similar to each other. To go into detail, the Euclidean distance from each observation in the data set to the centroids of the training years is shown in Figure 9.

Both series presented in Figure 9 are very similar. The period of time from 2007 to 2011, approximately, stands out due to its smaller range of values. The explanation for this fact is that the Water Level variable, which has great importance for the models of both pendulums, takes values within a less disperse range. For this reason, the distance is smaller since there are more points within this range of values (Figure 2).

On the other hand, the largest distances are found in 2016 and 2017, which are the farthest periods from the rest of the centroids in the graph Figure 8.

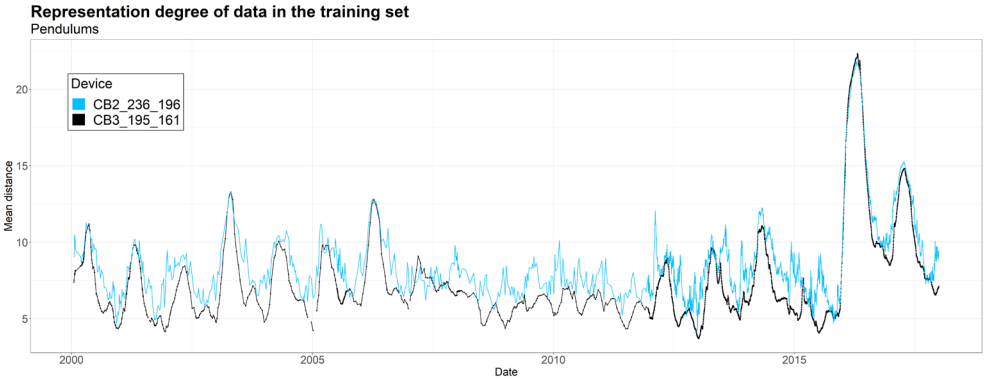

Figure 9. Series of the degree of representation, calculated as the Euclidean distance of the points of the different years to the centroids of the years used for training the models.

Figure 10 shows that, on average, the external conditions of 2016 and 2017 are the most different compared to other years. This is due to their low water level values. They are followed by the years 2006, 2003, and 2014, for both devices.

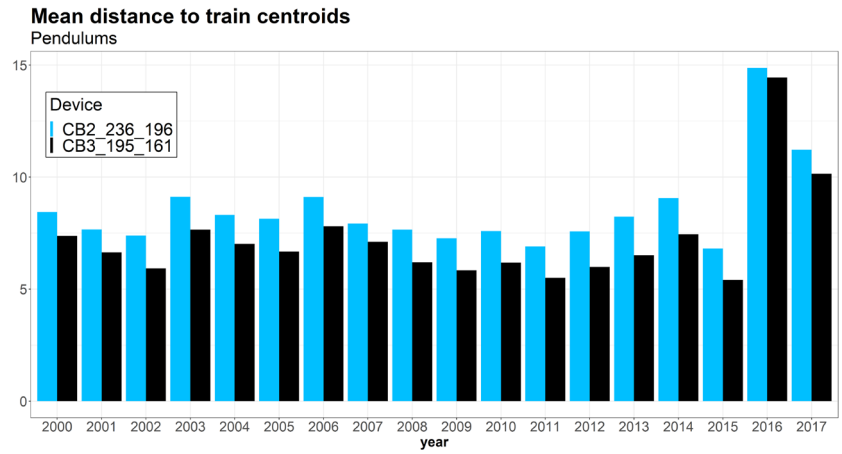

Figure 10. Mean Euclidean distance from the points of each year to the centroids of the training years.

The results of this section indicate that the differences between the most important external conditions of each pendulum are related to low water levels and steep drawdowns. Given that 2016 and 2017 are revealed to be odd and Machine Learning predictive models learn from data, the prediction error could be higher in these years.

3.5 *Models*

Among the results obtained when training the different models shown in Table 1 we found differences depending on the type of device. The SVM model is the optimal model to predict the series of both pendulums; for the pore pressure measuring device, the best model is BRT because it has the lowest $RMSE_{CV}$ values, while for leakage and joint opening, the most accurate is NN.

Table 1. Results of the models for each device. $RMSE_{CV}$ is the estimated error during the Cross Validation process. $RMSE_{val}$ is the error made on the validation set (year 2012).

	Displacement (pendulums)				**Joint opening**		**Pore pressure**		**Leakage**			
Device:	CB2_236_196		CB3_195_161		C4_C5		PZCB2		PZCB3		Seepage	
Model	RMSE CV	RMSE val	RMSE CV	RMSE val	RMSE CV	RMSE val	RMSE CV	RMSE val	RMSE CV	RMSE val	RMSE CV	RMSE val
SVM	1.794	1.771	0.409	0.334	0.25	0.232	0.759	0.485	0.479	0.633	3.106	2.609
BRT	2.334	2.562	0.554	0.395	0.277	0.215	0.56	0.609	0.349	0.586	3.143	3.231
NN	4.165	7.136	0.893	0.441	0.189	0.107	1.102	0.907	0.926	2.114	3.054	2.821
RF	2.747	3.103	0.62	0.641	0.333	0.312	0.763	1.43	0.479	0.514	3.235	3.092
HST	2.869	4.099	0.594	0.622	0.305	0.353	1.029	1.526	0.724	1.291	3.113	2.632
RRBB	3.74	3.029	0.803	0.428	0.653	0.359	1.91	2.314	0.853	1.565	3.635	2.908
LM	3.642	2.862	0.686	0.465	0.623	0.354	1.853	2.288	0.77	1.255	3.514	2.727

In some cases, the validation error is lower with other models than those mentioned in the previous paragraph, as in the case of pore pressure and leakage devices. However, as explained

in the Methodology section, the $RMSE_{CV}$ is a more robust estimator of the error because it uses more years in its calculation.

Figure 11 and Figure 12 show the results of the calibrated predictions during the CV and over the validation set of both pendulums. The series are significantly close to the actual values of the series. SVM seems to make a larger error in the high and low peaks of the years 2002, 2003, 2004 and 2005 in the case of the CB3_195_161 pendulum (Figure 11).

The short- and long-term predictions of both pendulums are also shown in these figures. The series corresponding to the CB3_195_161 pendulum appears to have a decreasing trend from approximately 2008 onwards, probably due to the trend of the water level during those years. From 2014 onward, the trend seems to disappear. The predictions for 2017 are within a narrower range than usual due to the large drop in the 2016 water level discussed in the exploratory analysis section that causes the 2017 water level to have low values (Figure 2). The same trend is observed in the series of predictions of the CB2_236_196 pendulum.

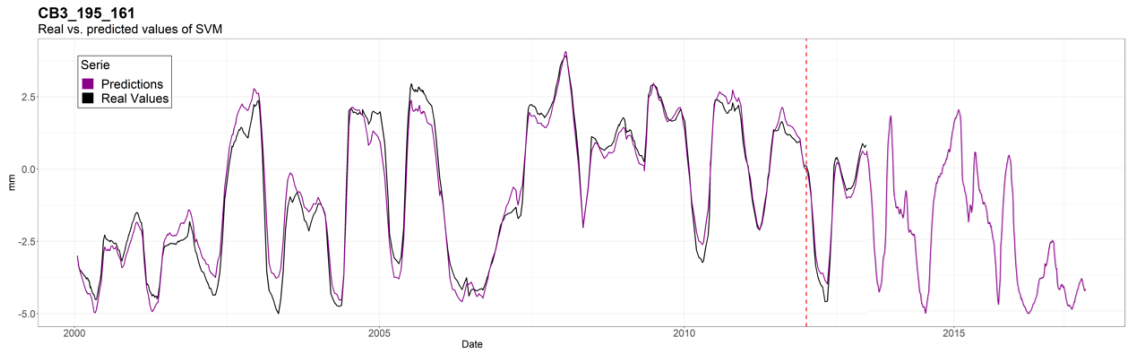

Figure 11. Series of real values of the CB3_195_161 pendulum and the predictions generated with the SVM model. The red dashed line separates the dates used to train the model and the validation set.

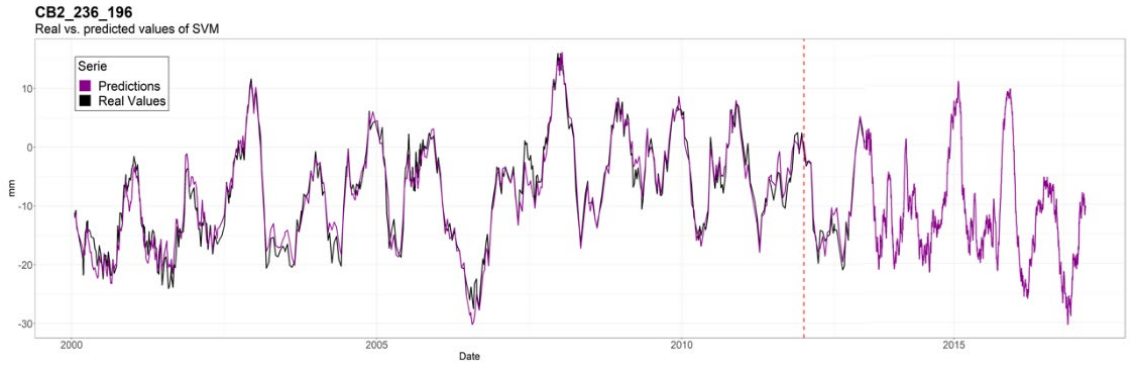

Figure 12. Series of real values of the CB2_236_196 pendulum and the predictions generated with the SVM model. The red dashed line separates the dates used to train the model and the validation set.

The outlier behavior of the water level in 2016 and 2017 makes it possible for the accuracy of the model to drop in those periods because that behavior has never been seen and the relationships between external conditions might not match those that the model has identified and configured.

3.6 *Warning Levels*

This section presents the results of the warning levels obtained by applying the formula explained in the Methodology chapter for each device.

Table 2. Table with the relevant information for the calculation of the warning levels of each target variable. Pto = c*σe.

	Displacement (pendulums)	**Joint opening**	**Pore pressure**	**Leakage**

Device:	CB2_236_196		CB3_195_161		C4_C5		PZCB2		PZCB3		Seepage	
c	**σ_e**	**Pto**	**σ_e**	**Pto**	**σ_e**	**Pto**	**σ_e**	**Pto**	**σ_e**	**Pto**	**σ_e**	**Pto**
1.5	1.851	2.777	0.454	0.681	0.198	0.296	0.582	0.873	0.357	0.535	3.096	4.645
2		3.703		0.909		0.395		1.163		0.713		6.193
3		5.554		1.363		0.593		1.745		1.070		9.289
4		7.405		1.817		0.790		2.327		1.427		12.385

The coefficient selected to determine the warning levels is 2, so the band of each instance will be its predicted value plus 2 times the standard deviation of the error.

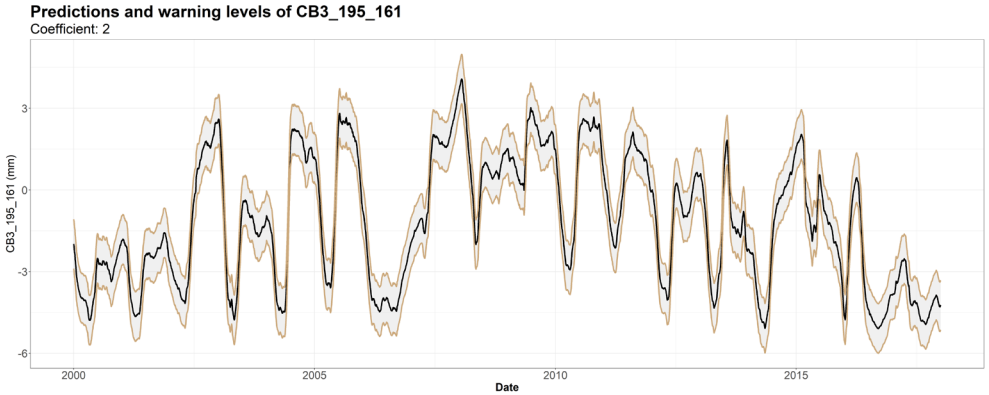

Figure 13. Warning levels of the pendulum CB3_195_161 calculated with c = 2.

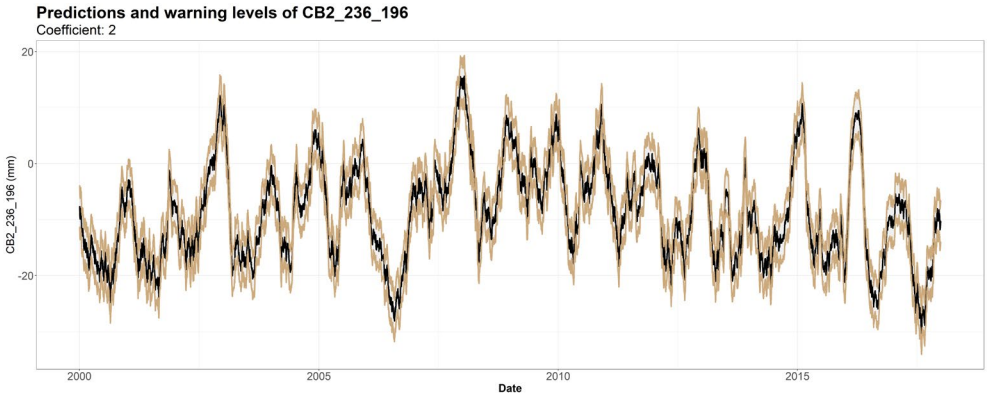

Figure 14. Warning levels of the pendulum CB2_236_196 calculated with c = 2.

4 CONCLUSSIONS

We have presented a methodology for the prediction of dam behavior through the selection of the most important variables and the optimal choice among models of different natures. Warning levels have been established based on the error of the best model. Moreover, an exploratory analysis of the external factors has been carried out, and we have gone into detail about the relationships between them and the pendulums in order to group the years according to the similarity between their external conditions.

The exploratory analysis showed different behaviors in the water level that are directly related to the behavior of the pendulum series. These changes in the series are linked to the water level drops observed in some years, such as 2003, 2006, and 20016.

On the other hand, the external factor that has the highest correlation with the pendulum series is the water level, more specifically the short-term moving averages in the case of pendulum CB3_195_161 and the long-term ones in the case of pendulum CB2_236_196. These synthetic variables are also the most important variables in both cases regarding predictive capacity, respectively.

Through our methodology we also identified groups of years based on the centroids and the distance series, or degree of similarity, of the external conditions of each year. The observed groups are similar to those intuited in the exploratory analysis: group 1, formed by 2000, 2001, 2003, 2003, 2004, 2006, 2006, 2012, 2013, 2013, 2014, 2015; group 2, consisting of 2005 and 2007 through 2011; and group 3, 2016 and 2017. On average, the most atypical years are 2016 and 2017, due to the steep drop in water level in 2016. This raises the possibility of an increase in our model error in those years.

The results on the optimal model for predicting dam behavior depend on the type of measurement involved. For dam displacement, the best model was SVM, for leakage and joint opening it was NN, while BRT gave the best results for pore pressure. The short- and long-term predictions of these models have a decreasing trend due to the observed decreasing trend of water level.

In summary, significantly accurate models have been built through the selection of the most important variables and the application of different algorithms, where the water level is highly correlated with the behavior of the dam.