Prediction and interpretation of dam response with boosted regression trees

Salazar, F. [1,2], Irazábal, J.[1] and Vicente, D.J.[1]

[1]CIMNE International Center for Numerical Methods in Engineering, Madrid, SPAIN
[2]Universitat Politècnica de Catalunya, Barcelona, SPAIN

E-mail: fsalazar@cimne.upc.edu

ABSTRACT: Predictive models based on boosted regression trees were fitted for computing the response of an arch dam in terms of radial displacements, joint opening, piezometric levels and seepage as a function of time series of external variables: water level, air temperature, rainfall and time. A generic procedure was followed for all outputs, supported by two software tools developed by the authors. Warning levels were generated based on the residuals. The analysis of the models showed the effect of the main loads, the thermal inertia for radial displacements, and changes over time for piezometric levels due to the cleaning of the drainage system performed in 2008.

1 Introduction

This document describes the process followed in response to Theme A proposed in the frame of the 16th International Benchmark Workshop on Numerical Analysis of Dams organized by the International Commission on Large Dams (ICOLD). The text focuses on the methods and tools used. Details about the proposed problem can be consulted in the description of the Theme and are therefore omitted here. Both the predictions and the interpretation were generated by means of two software tools previously developed by the authors for data visualization and preprocess[1] and for fitting models based on machine learning (ML)[2].

2 Methods

2.1 Preprocessing

Among the three versions of the starting data, we chose the file "ThemeA_data_fmt03.xlsx", which includes a common time vector for all variables and one record for each day in the period. For those variables with more than one value per day, the data set includes the mean. We checked such operation and the completeness of the time series. All preprocessing operations were performed using the free online tool "PREDATOR" developed by the authors [1].

We identified some missing values in the time series of water level, which were filled by linear interpolation (Figure 1). We verified that linear interpolation was reasonable, since the missing values were isolated. Time series for rainfall and temperatures featured no missing values. Since the entire upstream face of the dam is exposed for all values of water level below 196 m.a.s.l., we created a modified variable, in which all water levels lower than 196 are replaced by 196. It was called "modWL". Our approach includes generating derived variables:

- Moving averages of 7, 14, 30, 60 and 90 days for Water Level, modWL, T_a and T_b
- Cumulative sum of 7, 14, 30 and 60 days for rainfall.

---

[1] PREDATOR. https://cimnetest.shinyapps.io/PREDATOR/
[2] SOLDIER. https://cimnetest.shinyapps.io/SOLDIER/

This strategy allows for capturing delayed effects, such as the thermal inertia of concrete dams, as verified in previous works [3]. Two additional variables are automatically created by PREDATOR: “Year” and “month”. Only the first was used, to account for the time effect.

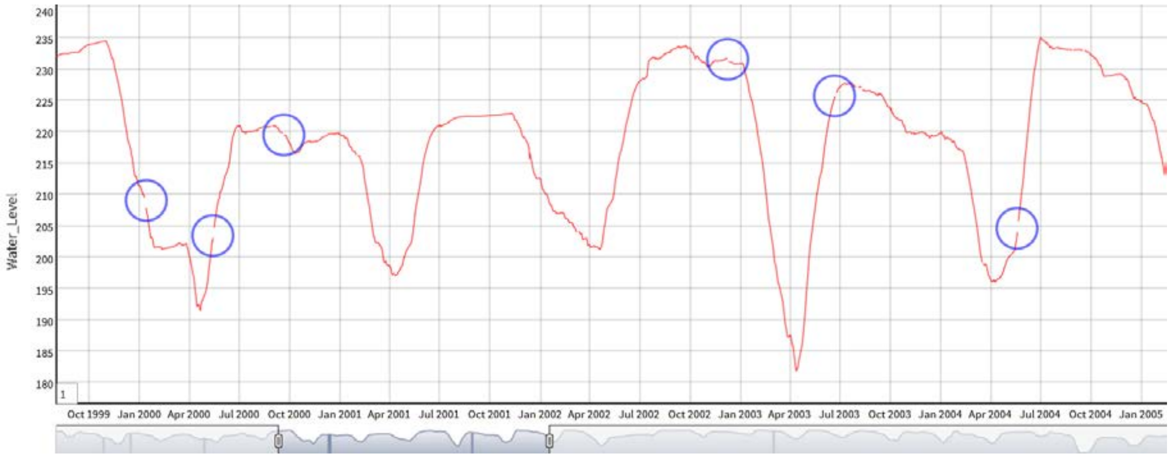

Figure 1: Some of the missing values in the time series of Water_Level. They were all filled by linear interpolation with PREDATOR.

Since the records for the period 2013-2017 are not available for the output variables, the training data set includes the period 1/1/2000-31/12/12. We saved a data file with that period and a separated one to generate the predictions, which only includes the input variables.

2.2 Model fitting

2.2.1. General approach

Our predictive models make use of the algorithm “Boosted Regression Trees” (BRT from now on). It is an ensemble method, widely applied in different fields, whose theoretical fundamentals can be found in the literature (e.g. [6]). We chose this algorithm on the basis of the conclusions of a previous comparative study among some of the most powerful ML algorithms [2], which were evaluated in terms of their accuracy for predicting dam behavior and their easiness of calibration and implementation. Further analysis showed its capability for identifying the effect of the loads on the response of the dam [3] and for detecting anomalies [4]. The algorithm is implemented in a free online application for fitting models for dam prediction called SOLDIER [1].

BRT models are highly flexible, i.e., they deduce the underlying relation between inputs and response from the training data without the need for a detailed selection of input variables or parameter calibration. This implies that a common process can be followed for generating models for predicting outputs of different nature, as is the case (displacements, joint opening, piezometric levels and seepage). The addition of irrelevant inputs has a minor effect on the predictions of the model. Nonetheless, for this particular case, we included variable selection in the process, as described below.

The application used allows for easily modifying the training and test periods, the input variables and the BRT parameters: bag fraction, interaction depth, number of trees and shrinkage. Although their effect on the predictive model is moderate, we followed a calibration process for each output.

In addition to the prediction of dam behavior, the organizers also asked for warning thresholds. We chose to define them as a function of the model accuracy, as recommended by ICOLD [5]. Therefore, a reliable estimate of the predictive accuracy is necessary for each model. BRTs always overfit to some extent, so computing model accuracy can be tricky. We chose fitting models using the period 2000-2010 and evaluating their accuracy on data for the most recent period (2011-2012). We verified that water level fluctuated along a relevant range in those last two years, which implies that the performance of the models for that period is sufficiently representative of their general prediction accuracy.

Fitting a BRT model for the size of the data sets used in the BW typically requires some seconds. However, the amount of possible combinations of input variables and model parameters is very high. The preprocessed input data includes 32 variables (original and derived variables), which means that the amount of possible combinations of inputs is 2^{32}-1=4.29x10^9. If each model took 10 seconds for training, considering one model for each set of inputs would require 1.98x10^5 days. Such process is therefore unfeasible. Instead, we followed a process for variable selection and model calibration that includes the following steps:

1. Interactively try options for each output using SOLDIER and visualize results. The options include both the input variables and the BRT parameters. In view of the results, make decisions to reduce the amount of possible combinations of inputs and model parameters to analyze.
2. Select a feasible set of combinations and perform random search model calibration.
3. Visually verify the candidate models –those with lower predictive error– back with SOLDIER

2.2.2. Preliminary interactive exploration

The first stage involved the following steps:

1. Fit a model with default training parameters, the period 1/1/2000-31/12/11 and all inputs: rainfall, temperature, water level –and their corresponding derived variables– and Year.
2. Interactively evaluate the effect of the model parameters: interaction depth, shrinkage and number of trees. The goodness of fit is evaluated by means of the mean absolute error (MAE).
3. Check the effect of the input variables with the relative influence and the partial dependence plots.
4. Check the residual distribution and evolution with time.

Figure 2 shows an example screenshot of the interface showing the mentioned information.

As a result of this process, the overall options were narrowed and the following decisions were made:

- T_a and its derived variables were discarded, since T_b systematically resulted in higher relevance, i.e., stronger association with the responses.
- modWL and its moving averages were chosen instead of the raw Water_Level. In this case, the difference was moderate.
- Rainfall was neglected, since had negligible effect in all cases.

Figure 2: Interface of SOLDIER during model fitting. Target and input variable selection (top left), training period and model parameters (bottom left), accuracy measures for training and test (top right) and graphical representation of predictions, observations and residuals (bottom right)

2.2.3. Random search calibration

This step was performed by means of ad-hoc scripts written in the R programming language. For each output, the same process was followed:

1. 100 combination of inputs were considered with the following criteria:
   - Time was taken as input in half of the models, and excluded in the others.
   - modWL and T_b_14 were always taken as inputs.
   - A random subset of the remaining inputs was taken to complete the input set.
2. For each set of inputs, all possible combinations of model parameters included in Table 1 were used, i.e., 36 versions of the model
3. As before, models were fitted over the period 2000-2010 and their performance was assessed for 2011-2012. MAE was computed both for the training and the test sets.
4. The models were evaluated on the basis of a score computed with Equation (1), and that with lower value was selected. With this criterion, when several models were obtained with similar precision in the test set, the one with the highest train error and therefore the lowest risk of overfitting was favored.

$$Score = MAE_{test} + (MAE_{test} - MAE_{train}) \quad (1)$$

Table 1: BRT Model parameters considered for each combination of inputs

Parameter	Values
Number of trees	1000, 2000, 3000
Shrinkage	0.01, 0.005, 0.002
Interaction depth	2, 3, 4, 5

2.2.4. Final model selection

The models selected in the previous step, i.e., those with lower score for every output, were again loaded in SOLDIER and verified: the accuracy for training and testing, the residual distribution and its evolution in time, and the importance of the variables.

In addition to the visual verification of model accuracy, this step allowed for checking that no spurious effects of any input variable were considered. We put special attention on the effect of time, which encompasses the information not recorded by the input variables available, and which is the input most prone to overfit. In particular, when the best model excluded time as input, the final check involved comparing the results with those obtained with a modified model adding "Year" to the input set.

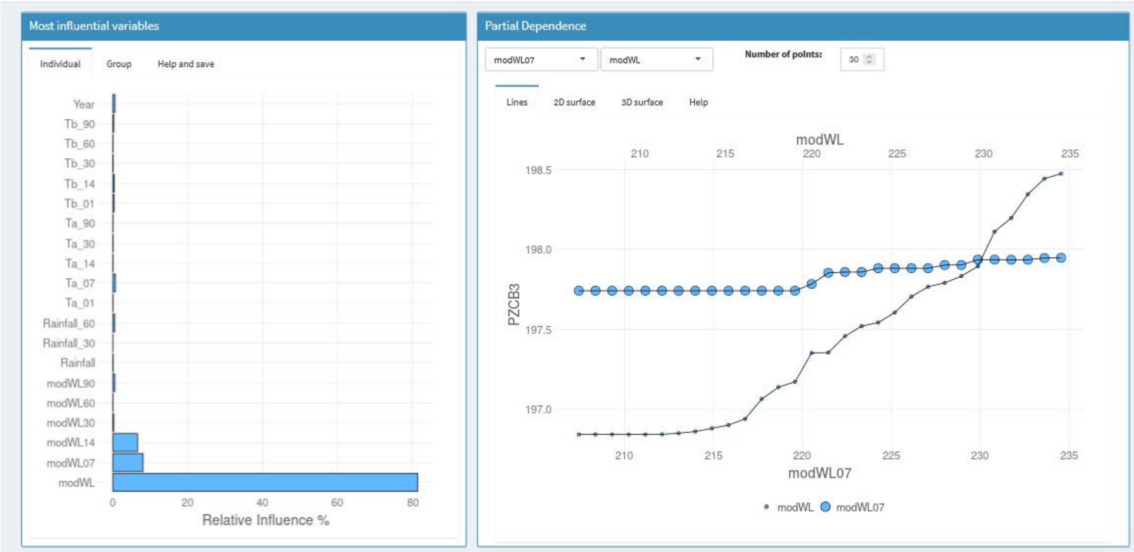

Figure 3: Interface of SOLDIER for model interpretation. Classification of inputs as a function of their relevance on the response (left) and partial effect of the most important inputs (right).

2.3 Generation of predictions and warning levels

The final models were loaded back in R with another specific script, together with the test data for generating predictions. The density functions of the residuals were generated for years 2011 and 2012 to check normality. The main statistics of the residuals were computed: mean, median, standard deviation and quantiles for 0% and 100%. They were all plotted over the histograms of residuals. Although residuals followed distributions close to normal for many outputs, this was not the case for seepage. We finally decided to take the quantiles for computing the prediction intervals.

We also corrected the bias in predictions by adding the median of the residuals. As a result, predictions and warning thresholds were generated as described in Equations (2), (3) and (4).

$$Res = Obs - Pred \;\rightarrow Obs = Pred + Res \;\rightarrow Pred_{corr} = Pred + median(Res) \qquad (2)$$

$$Upp = \; Pred + Res_{q100} \qquad (3)$$

$$Low = \; Pred + Res_{q0} \qquad (4)$$

Where Res = residuals; Obs = Observations; $Pred$ = Predictions; $Pred_{corr}$ = corrected predictions; Upp = upper warning threshold; Low = lower warning threshold; Res_{q100} = Quantile 100 of residuals; Res_{q0} = Quantile 0 of residuals.

3 Results and discussion

Since the predictions are the main outcome of the analysis, and they will be evaluated by the organizers, only the most relevant aspects of model interpretation are described in this section for each output.

3.1 Displacements

3.1.1. CB2_236_196

The calibrated model included only three inputs: modWL, T_b_01 and T_b_14. The effect of the inputs on the displacement agrees with engineering knowledge, i.e., high water level and low temperature are associated with higher deformation in the downstream direction, and vice versa. The moving average of 14 days or air temperature has more influence than the daily temperature, which shows the thermal inertia of the dam.

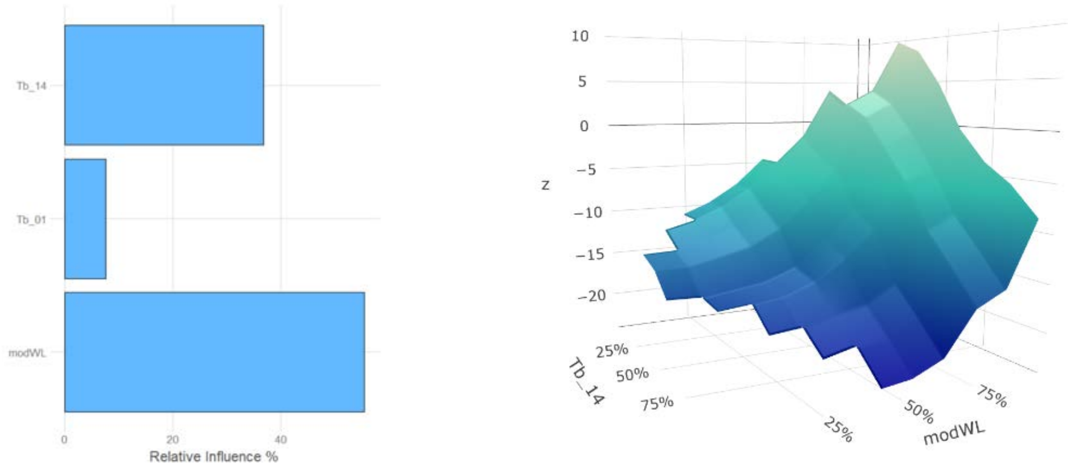

Figure 4: Left: relative influence of the selected inputs. Right: combined partial effect of hydrostatic load (modWL) and air temperature (T_b_14) on CB2_236_196

3.1.2. CB3_195_161

The final model for the displacement in the foundation is also based on three inputs, but T_b_60 is taken instead of T_b_01. Nonetheless, the effect of air temperature is much lower than before, as can be expected, since the foundation is less exposed to the variations in ambient temperature. As before, adding time as input resulted in similar accuracy.

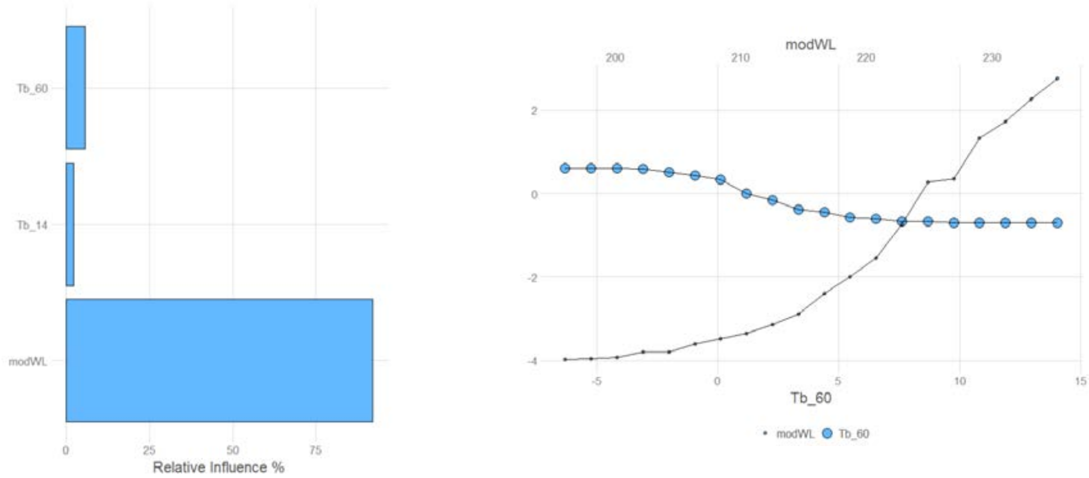

Figure 5: Left: relative influence of the selected inputs. Right: partial dependence of the displacements at the foundation on the water level (modWL) and air temperature (Tb_60).

3.2 Joint opening

The final set of inputs for joint opening includes several moving averages of both main loads. In particular, modWL, modWL60, modWL90, TB_07, TB_14, TB_30, TB_90. In this case, adding time as inputs led to lower accuracy for the test set (2011-2012), which means that induced overfitting.

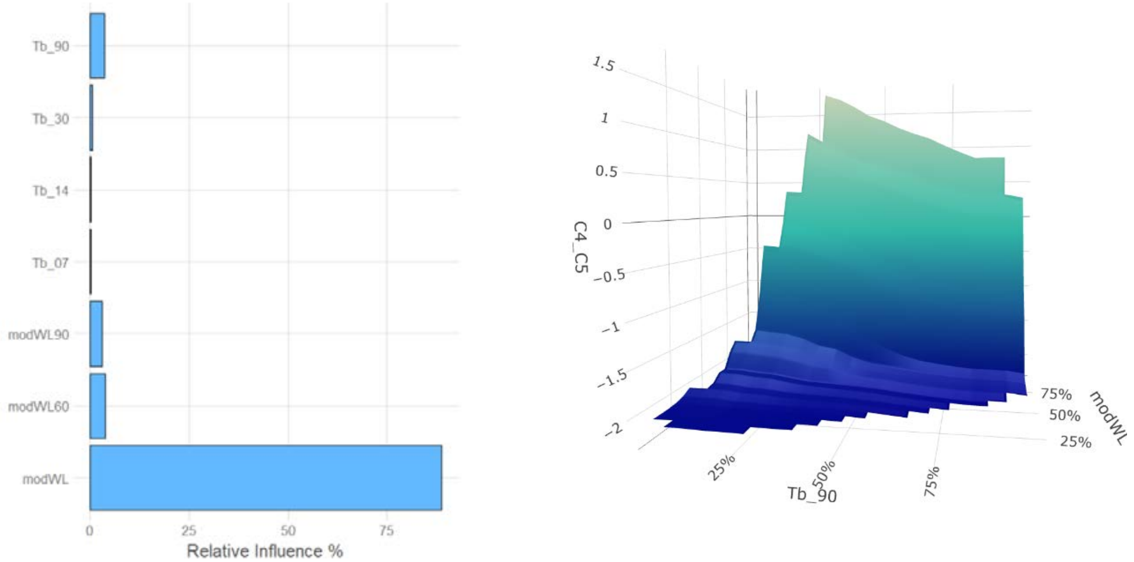

Figure 6: Left: relative influence of the selected inputs on joint opening. Instantaneous water level is by far the most important load. As for the air temperature, the effect increases with the period of the moving average. Right: combined effect of modWL and Tb_90 shows that the influence of air temperature is more important for high water level.

3.3 Piezometric levels

3.3.1. PZCB2

The calibration process resulted in a model including time as input for PZCB2. This implies that the algorithm identified an evolution on the dam response, i.e., for a given combination of water level and temperature, the piezometric level changed over time (Figure 7).

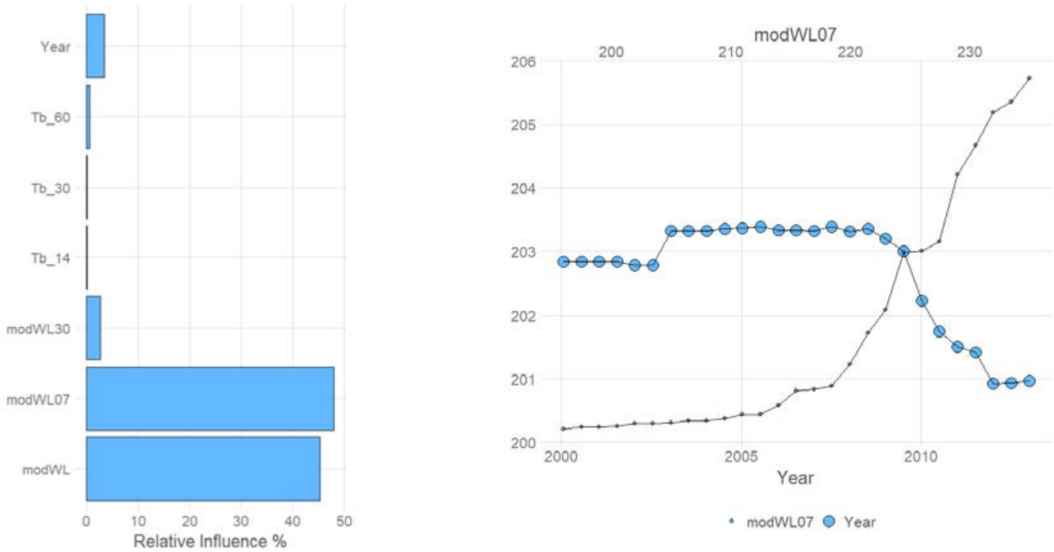

Figure 7: Left: relative influence of the selected inputs on PZCB2. Time (Year) has a relevant effect. Right: partial dependence on modWL and time. The algorithm identified a sharp decrease in PZCB2 on 2010 and stabilization in 2012.

This can be verified by exploring the time series of the measured data, included in Figure 11 in the Theme A description, as well as in the scatterplot in Figure 8.

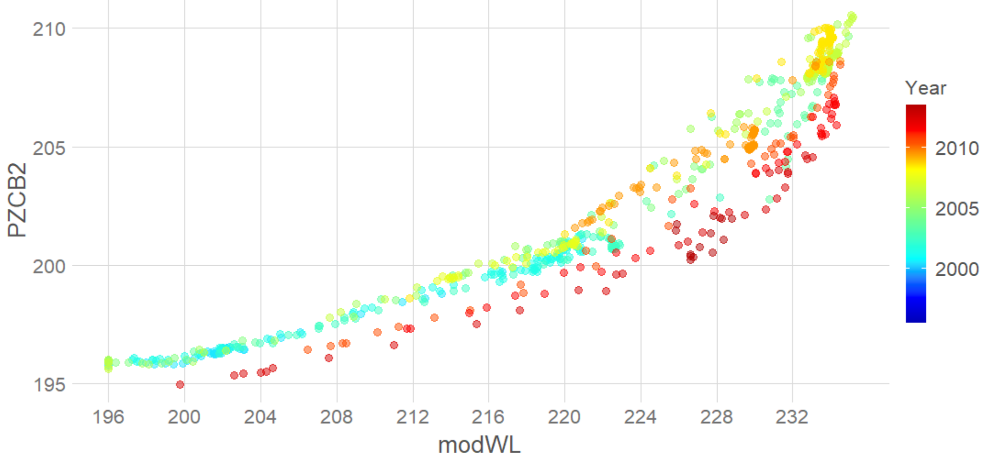

Figure 8. Scatterplot of PZCB2 as a function of water level, colored by time. Recent values are lower for a given hydrostatic load.

The final model included both modWL and modWL_07, being the latter even more influential. This may suggest some inertia in PZCB2 in response to changes in water level. However, we verified that changes in predictions were negligible after removing modWL_07. Therefore, the higher influence may be a spurious result due to the high similarity between both inputs.

3.3.2 PZCB3

The description of the Theme mentioned a change on PZCB3 in 2008, as well as a period with missing data. This change along time was also identified by the algorithm and “Year” was included in the input set resulting from the calibration process. The flexibility of BRTs allowed for using the same fitting process for this target, for which a clear change was known in advance. Indeed, the interpretation of the model (Figure 9 right) shows the mentioned change in 2008. Nonetheless, the algorithm also found another change in the last year of the period provided (2012), which is also observed in the exploratory plots (Figure 10).

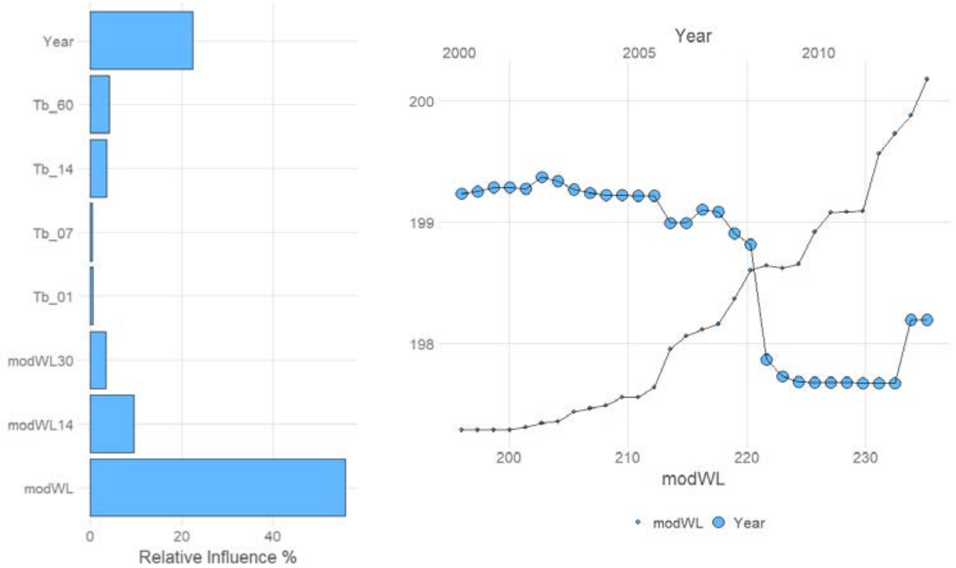

Figure 9: Left: relative influence of the selected inputs on PZCB3. Water level is the most important input, and its moving averages feature decreasing influence as the period increases. Time (Year) is considered as highly relevant. Right: partial dependence on the most relevant inputs. The effect of time shows the known change in 2008, but also a shift in 2012.

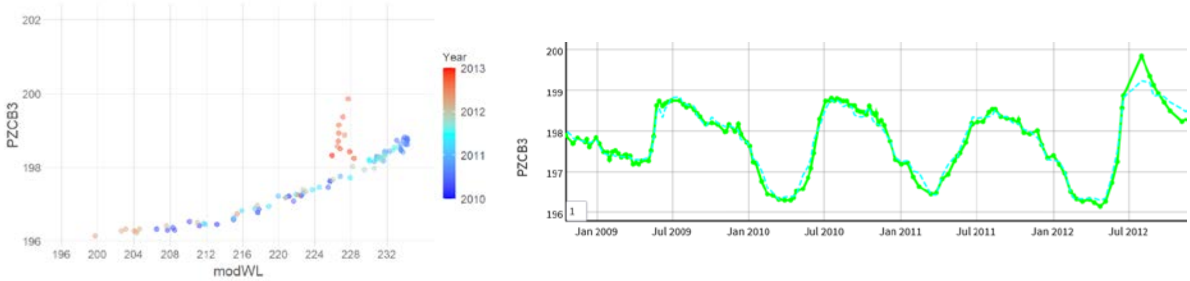

Figure 10. Verification of the abnormal records of PZCB3 in 2012 by exploration of the period 2010-2012. Left: scatterplot. Right: time series.

3.4 Seepage

Predictions of seepage were less accurate than those for the remaining targets. Although the MAE can be considered acceptable (around 2 l/sec), large errors occurred for high flows (Figure 11). Water level is clearly the more influential input. However, adding T_b consistently results in higher prediction accuracy. This may be due to some effect of temperature in joint opening and subsequently in higher leakage flows, but any conclusion is unreliable due to the low reading frequency.

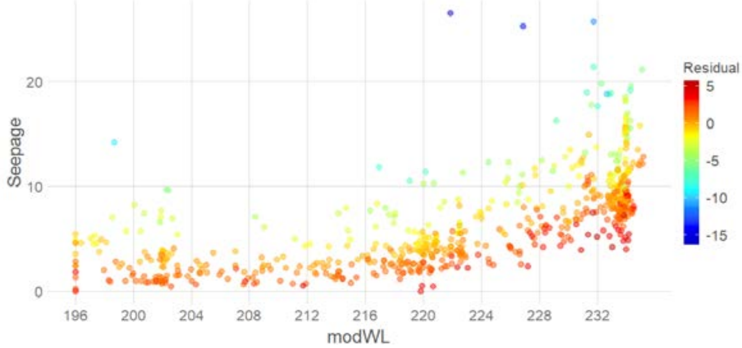

Figure 11. Scatterplot of Seepage as a function of water level, colored by residual (prediction error). Large errors are observed for some high flows.

Although the time series of seepage is noisier than the others, the formulators mentioned no anomalies for this variable. As a result, no record was discarded even though some look like outliers. For instance, in the period Dec/2008-Mar/2009, seepage gently decreases, apparently following the evolution of water level. However, two values higher than 25 l/sec are registered, clearly out of the overall trend (Figure 12).

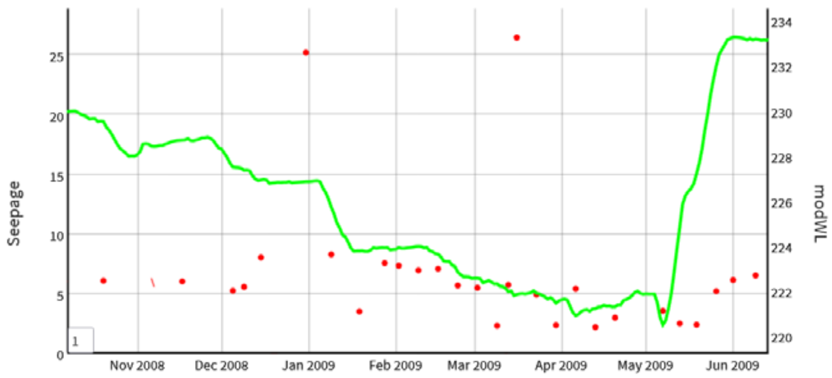

Figure 12. Time series of seepage (red dots) and water level (green line) from Dec/2008 to Mar/2009. Two records of seepage seem to be out of the general trend.

4 Conclusions

Predictive models based on ML (BRTs) for response variables of different nature were generated and analyzed with a general process supported by two software tools. The flexibility of BRT models allowed for performing all posed tasks with minor changes. For piezometric levels, the entire available period was used without the need to include any modification to account for the known change in behavior in 2008 due to the cleaning performed on the drainage system. On the contrary, it was automatically identified by the model.

The interpretation of the models showed the effect of the main loads generally in accordance with engineering knowledge: high hydrostatic loads are associated to displacement in the downstream direction, high seepage, joint opening and piezometric levels; time effect and thermal inertia were identified for CB2_236_196.

We generated warning levels based on the quantiles of the residuals for the period 2011-2012. Although water level followed a similar pattern in those years to that observed in 2000-2009, such levels may not be useful for the entire prediction period because: a) water level was abnormally low in 2016 and 2017; b) we recommend updating the models every year to include additional information and possible changes in dam response. In any case, a record above the upper limit –or below the lower– should be taken as a warning for additional actions to make before issuing an alarm. These may include verification of related targets, increase of reading frequency and follow up of the evolution of the abnormal variable.

5 Acknowledgements

The authors acknowledge the contribution of André Conde Vázquez to the development of PREDATOR and SOLDIER.

This work was partially funded by the Spanish Ministry of Science, Innovation and Universities through the Project TRISTAN (RTI2018-094785-B-I00). The authors also acknowledge financial support from the Spanish Ministry of Economy and Competitiveness, through the "Severo Ochoa Programme for Centres of Excellence in R & D" (CEX2018-000797-S), and from the Generalitat de Catalunya through the CERCA Program.

6 References

[1] Salazar, F., Kohler, R., Conde, A., & Landstorfer, F. (2019, September). Interpretation of dam monitoring data combining visualisation tools and machine learning. Eberlaste Dam Case Study. In *ICOLD International Benchmark Workshop on Numerical Analysis of Dams, ICOLD-BW* (pp. 863-874). Springer, Cham.

[2] Salazar, F., Toledo, M. A., Oñate, E., & Morán, R. (2015). An empirical comparison of machine learning techniques for dam behaviour modelling. *Structural Safety*, *56*, 9-17.

[3] Salazar, F., Toledo, M. Á., Oñate, E., & Suárez, B. (2016). Interpretation of dam deformation and leakage with boosted regression trees. *Engineering Structures*, *119*, 230-251.

[4] Salazar, F., Toledo, M. Á., González, J. M., & Oñate, E. (2017). Early detection of anomalies in dam performance: A methodology based on boosted regression trees. *Structural Control and Health Monitoring*, *24*(11), e2012.

[5] ICOLD. (2018). *Dam Surveillance Guide* (Bulletin nº 158). CRC Press.

[6] Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of animal ecology*, *77*(4), 802-813.