Early detection of anomalies in dam performance: A methodology based on boosted regression trees

Fernando Salazar1 | Miguel Ángel Toledo2 | José Manuel González1 | Eugenio Oñate1

1International Center for Numerical Methods in Engineering (CIMNE), Campus Norte UPC, Gran Capitán S/N Barcelona, 08034, Spain
2Department of Civil Engineering: Hydraulics, Energy and Environment, Technical University of Madrid (UPM), Profesor Aranguren S/N Madrid, 28040, Spain

Correspondence
Fernando Salazar, International Center for Numerical Methods in Engineering (CIMNE), Campus Norte UPC, Gran Capitán S/N, Barcelona, 08034, Spain. Email: fsalazar@cimne.upc.edu

Funding information
Spanish Ministry of Economy and Competitiveness (Ministerio de Economía y Competitividad, MINECO), Grant/Award Number: IPT-2012-0813-390000, BIA2013-49018-C2-1-R and BIA2013-49018-C2-2-R

Summary
The advances in information and communication technologies led to a general trend towards the availability of more detailed information on dam behaviour. This allows applying advanced data-based algorithms in its analysis, which has been reflected in an increasing interest in the field. However, most of the related literature is limited to the evaluation of model prediction accuracy, whereas the ulterior objective of data analysis is dam safety assessment. In this work, a machine-learning algorithm (boosted regression trees) is the core of a methodology for early detection of anomalies. It also includes a criterion to determine whether certain discrepancy between predictions and observations is normal, a procedure to compute a realistic estimate of the model accuracy, and an original approach to identify extraordinary load combinations. The performance of causal and noncausal models is assessed in terms of their ability to detect different types of anomalies, which were artificially introduced on reference time series generated with a numerical model of a 100-m-high arch dam. The final approach was implemented in an online application to visualise the results in an intuitive way to support decision making.

KEYWORDS
anomaly detection, boosted regression trees, dam monitoring, dam safety, machine learning

1 | INTRODUCTION

Dam safety is an area of growing interest: Our societies demand increasing safety levels, and the average age of dams is high in many countries, which increases the need for control and maintenance operations. The advances in information and communication technologies led to relevant improvements in the performance of monitoring systems, in terms of both accuracy and reliability of the devices, security of the communications, and reading frequency. All this resulted in more information available on the behaviour of the structure.1

This increase in the amount of available data led to the use of more powerful tools for its analysis, from enhanced versions of the multiple linear regression (e.g., Mata et al2) up to algorithms developed in the field of machine learning, such as neural networks (NN),3 support vector machines,4,5 or adaptive neuro-fuzzy inference systems,6 among others.7,8

However, these methods are still not widely applied by practitioners, who mostly limit the data analysis to graphical exploration of the time series of data,9 along with simple statistical models.1,10

The vast majority of examples of application of advanced tools focus on the development of behaviour models to predict the value of a given response variable of the dam (e.g., radial displacement) as a function of the loads. The prediction is compared to the actually observed data, and some error index is computed. In most cases, the results are more accurate than those obtained by conventional methods (e.g., Mata).3

In general, these techniques offered some advantages over conventional statistical methods, in terms of greater accuracy, flexibility, or ability to interpret dam behaviour.11
However, the main objective of dam safety is to prevent failures, for which anomalies need to be detected at early stage. The capability of predictive models to identify anomalies has been much less frequently studied. Mata et al. developed a model based on linear discriminant analysis for the early detection of developing failure scenarios. This methodology belongs to the Type 2 among those defined by Hodge and Austin. The system is trained with both normal and abnormal behaviour data and classifies new inputs as belonging to one of those categories. The drawback of this approach is that the failure mode must be defined beforehand and simulated with sufficient accuracy to provide the training data. Hence, the system is specific for the failure mode considered.

Jung et al. used a similar approach: Abnormal situations were defined based on the discrepancy between the model predictions and the observed data. This method focuses on embankment dam piezometer data, and only the reservoir level is considered as external variable (although they acknowledge that the rainfall can also be influential). It is not clear whether this methodology could be applied to other dam typologies or response variables.

Cheng and Zeng presented a methodology based on the definition of some control limits, which depends on the prediction error of a regression model. In addition, they proposed a classification of anomalies based on the trend of the deviation and on how the overall deviation is distributed among the devices considered. It has the advantage of being simultaneously applied to a set of devices, although the case study presented is simple and the test period considered very short (30 days), as compared to the available data (1,555 days).

Other examples of application of advanced tools together with prediction intervals have been published by Gamse and Oberguggenberger, who employed the procedure of probabilistic quality control, Yu et al. based on principal component analysis, Kao and Loh, who used principal component analysis together with NN, Li et al., who considered the autocorrelation of the residuals, and Loh et al., who presented models for short- and long-term prediction.

Most of these works follow a conceptually similar methodology: A prediction model is built; the density function of the residuals is calculated and used to define the prediction intervals, which are applied to detect anomalies. In all cases, the efficiency is verified by means of its application to a short period of records. As an exception, Jung et al. and Mata et al. used abnormal data obtained from finite element models (FEM).

The main differences among authors lie in the prediction method used (parametric or nonparametric, static or autoregressive, etc.). In this article, a similar methodology is presented, with some innovative features:

- The prediction model is based on boosted regression trees (BRTs), which showed to be more accurate than other machine learning and statistical tools in previous works.
- Causal, noncausal, and autoregressive models are considered and jointly analysed.
- Artificially generated data are taken as reference. They were obtained from an FEM model considering the coupling between thermal and hydrostatic loads. This allows to identify normal and abnormal behaviour, as observed by some authors. In this work, the FEM results are compared to actually observed data to verify their reliability.
- A methodology is proposed to neglect false anomalies due to the occurrence of extraordinary loads. It is based on the values of the two main actions (thermal and hydrostatic).
- Three types of anomalies are considered, affecting both to isolated devices and to the whole structure.
- Although radial displacements in an arch dam were selected for the case study, the method can be applied to other dam typologies and response variables. Moreover, it adapts well to different amount and type of input variables, due to the great flexibility and robustness of BRTs.

The rest of the paper is organised as follows. A brief introduction to BRT is included, together with the main ingredients of the methodology. Then, the case study is described: the dam, the available data, the FEM model, and the artificial anomalies considered. Section 3 contains the results in terms of ability to detect different types of anomalies. The final version was implemented in an interactive tool, which is presented in the same section. Finally, overall conclusions are derived, and suggestions for practical application are provided.

2 METHODS

In previous studies, BRTs showed to be appropriate to build predictive models, mainly because of its high accuracy and flexibility. The algorithm was further analysed in terms of model interpretation, and it was verified that useful information can be drawn as regards dam performance. As a result, BRT was selected in this work as the predictive model, though the overall methodology may also be employed with other algorithms.

In what follows, \( Y \in \mathbb{R} \) stands for the output variable (radial displacement), which is estimated as a function of some inputs \( X \) (e.g., reservoir level and air temperature): \( Y \approx \hat{Y} = F(X) \). The observed values are denoted as \( (x_i, y_i), i = 1, \ldots, N \), where \( N \) is the number of observations. Each \( x_i \) is a vector with \( p \) components, each of which is referred to as \( x_{ij} \). Similarly, \( X^j, j = 1, \ldots, p \) stands for each dimension of the input space.
The outputs considered correspond to eight radial displacements in four plumb lines (two measurements per each plumb line). The employed notation and their location within the dam body are shown in Figure 7.

2.1 | Boosted regression trees

BRT models belong to the category of ensemble methods, because the prediction is based on the contribution of a number of simple models (weak or base learners). In particular, two algorithms are combined in BRTs: decision trees\(^ {20}\) for the base learners and boosting\(^ {21}\) for averaging. The fundamentals of both algorithms and that of BRTs can be found in the scientific literature,\(^ {22, 24}\) as well as in previous studies.\(^ {11}\) For the sake of completeness, a brief introduction follows.

Regression trees are based on the recursive division of the training space into disjointed regions. The prediction is generally the mean of the output variable for the observations within each region. They were first proposed by Breiman et al.\(^ {20}\) and underwent high degree of development from then on.

In the general case, when several inputs are considered, the best split point for each is calculated and that resulting in greater error reduction is chosen. This procedure allows for automatic selection of the most relevant predictors.

Regression trees are robust, require little data preprocessing, and can automatically reproduce nonlinear relations, as well as interaction among predictors. By contrast, they are unstable, that is, small variations in the training data may result in highly different results.\(^ {24}\)

Boosting is a general procedure to build ensemble predictive tools,\(^ {21}\) based on the combination of a number of simple models. The overall prediction is computed as a weighted sum of the output of each model in the ensemble. The rationale behind the method is that the average of the prediction of many simple learners can outperform that from a complex one.\(^ {25}\)

The main steps of the original boosting algorithm for regression trees and the squared error loss function can be summarised as follows:\(^ {26}\):

1. Start predicting with the average of the observations (constant):
   \[
   F_0(X) = f_0(X) = \bar{y}_i.
   \]
2. For \(m = 1\) to \(M\),
   (a) Compute the prediction error on the training set:
   \[
   \tilde{y}_i = y_i - F_{m-1}(x_i).
   \]
   (b) Draw a random subsample of the training set \(S_m\).
   (c) Consider \(S_m\) and fit a new regression tree to the residuals of the previous ensemble:
   \[
   \tilde{y}_i \approx f_m(X), t \in S_m.
   \]
   (d) Update the ensemble:
   \[
   F_m(X) \Leftarrow F_{m-1}(X) + f_m(X).
   \]
3. \(F_M\) is the final model.

A regularisation parameter \(\nu \in (0, 1)\) is typically added to avoid overfitting, so that step d turns into
\[
F_m(X) \Leftarrow F_{m-1}(X) + \nu \cdot f_m(X).
\]

Based on the results of previous studies,\(^ {8, 11}\) the models employed in this work initially contained 1,000 trees of two levels (four leaves), which were later pruned to the final shape via five-fold cross-validation. The regularisation parameter \(\nu\) was set to 0.01. All the calculations were performed in the R environment\(^ {27}\) with the \textit{gbm} library.\(^ {28}\)

2.2 | Prediction intervals

As mentioned above, most of the published works on the application of data-based models in dam monitoring are limited to the assessment of the model accuracy. However, the main practical utility of these models is the early detection of anomalies, for which it is necessary to compare the predictions with monitoring readings and verify whether they fall within a predefined range. If the residual density function follows a normal distribution, that range can be defined in terms of the standard deviation of the residuals. For example, Kao and Loh\(^ {17}\) presented the 99% prediction intervals for models based on NN, whereas Jung et al.\(^ {14}\) tested 1, 2, and 3 standard deviations of the residuals as the width of the prediction interval.

Based on previous studies with models based on BRTs,\(^ {29}\) the prediction interval in this work was set to \([\mu - 2 \cdot \text{sd}_{\text{res}}, \mu + 2 \cdot \text{sd}_{\text{res}}]\), being \(\mu\) and \(\text{sd}_{\text{res}}\) the mean and the standard deviation of the residuals, respectively. Special attention was paid to the determination of a realistic residual distribution. It is well known that the accuracy of a machine-learning prediction model must be calculated from a data set not used for model fitting\(^ {30}\) (validation set). In the case of time series, this validation set should be more recent in time than the training data, because in practice, the model is used for predicting a time period subsequent to the training data.\(^ {31}\)

The holdout cross-validation method meets this requirement, with the most recent data in the holdout set (Figure 1).

However, this implies discarding the most recent data for the model fit, which are generally the most useful, because
they represent the most similar behaviour to that to be predicted (assuming there may be a gradual change in behaviour over time). Moreover, the validation data may be biased, if they correspond, for instance, to a especially warm (or cold) period.

To overcome these drawbacks, while maintaining good estimate of the prediction error, an approach based on the holdout cross-validation method suggested by Arlot and Celisse\cite{Arlot:2010} for nonstationary time series data was employed.

The proposed method takes into account the following specific aspects of dam behaviour: (a) changes in the dam–foundation system are generally gradual and (b) dam behaviour models are typically revised annually, coinciding with the update of safety reports.

Let us consider that a behaviour model is to be fitted at the beginning of year $Z_i$, to be applied for anomaly detection during that year. The available data corresponds to the years $Z_1 \ldots Z_{i-1}$, with $Z_1$ being the initial year of dam operation. With the simple holdout method, a model is fitted with data in years $Z_1 \ldots Z_{i-2}$, whose accuracy is evaluated on data in $Z_{i-1}$.

In this work, a minimum training period of 5 years was considered. This value was chosen in view of (a) the results of previous studies\cite{SALAZAR:2023} and (b) the evolution of model accuracy on the reference data, as described in Section 3.2. Then, an iterative process is followed to reduce potential bias in the loads during $Z_{i-1}$. A set of predictions is generated as follows:

- For $k = 5 \ldots i - 2$
  - Fit a model $M_k$ trained with the period $Z_1 \ldots Z_k$.
  - Compute $R_k$ as the residuals of $M_k$ when predicting year $Z_{k+1}$.
  - Compute the mean ($\mu_k$) and standard deviation ($\text{sd}_{\text{res},k}$) of $R_k$.

At the end of the process, residuals for a set of models $M_k, k = 5 \ldots i - 2$ are obtained, with the particularity that they are computed over different time periods, always subsequent to the training set ($Z_6 \cdots Z_{i-1}$). That is, the amount of observations in the training sample increases and is used to predict the following year. The potential bias of some abnormal loads for 1 year is compensated by averaging, whereas a realistic prediction error is achieved, because it is always based on precedent data. A similar approach was employed by Herrera et al. to estimate demand in water supply networks, who employed the term growing window strategy\cite{Herrera:2013}.

Additionally, because the model accuracy typically increases as the training data grows, the actual model accuracy for the application period (year $Z_i$) will be more similar to that obtained for $Z_{i-1}$. Hence, $R_{i-2}$ is more representative of the expected model performance for $Z_i$. To account for this issue, the prediction intervals are based on a weighted average of $\mu_k$ and $\text{sd}_{\text{res},k}$. In particular, the weights for each year decrease geometrically from the most recent to the first available. A schematic representation of the procedure is included in Figure 2.

```
\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure2.png}
\caption{Graphical representation of the weighted growing window cross-validation procedure. The prediction interval is estimated as a function of the weighted average of the standard deviation of the residuals for previous years; each one is computed from a model trained with a different training set.}
\end{figure}
```

Finally, to take advantage of all the available data, a model is fitted with the entire period $Z_1 \ldots Z_{i-1}$, with which the predictions for the following year ($Z_i$) are computed.

Since the test set becomes part of the validation period in the subsequent years, the residuals generated during the application of the model in the test period can be added to those computed for previous years, so that there is no need to repeat the whole process: The previous residuals can be employed to obtain the new prediction interval, after updating the correspondent weights.

### 2.3 Causal and noncausal models

BRT models are robust against the presence of uninformative or highly correlated predictors.\cite{SALAZAR:2023} Hence, variable selection is much less influential for tree-based methods than for other machine-learning tools.\cite{SALAZAR:2023}
In the vast majority of published works, variables correspondent to environmental actions are considered as predictors: air temperature and hydrostatic load. Also, a time-dependent term is typically included, to identify possible variations in dam behaviour over the period of analysis. This criterion leads to causal models, because there is some causality relation between each input and the dam response, which can be identified by means of model interpretation.[11] This approach was also followed in this work to build the causal models, though several variables derived from those actually measured at the dam site (reservoir level and the average daily temperature) were also included. They are listed in Table 1. A priori, a model of this type is expected to detect reading errors and changes in dam behaviour. However, its accuracy might be improved, because the response of the dam may depend on variables not available, such as the maximum and minimum daily temperatures or the solar radiation.

A more accurate model can be obtained by adding dam response variables to the set of inputs. This means that each radial displacement is included in the input set to predict other radial displacements. This version will in principle give greater precision, because the record from a neighbouring device (e.g., another station of the same pendulum) implicitly contains the effect of external variables not considered in the causal version. By contrast, this model might not be able to detect anomalies affecting several devices. For example, a slide in a block of a concrete gravity dam will be reflected in all stations of the correspondent plumb line; therefore, the relation between the hydrostatic load and the displacement would be abnormal, whereas the relationship between several readings of the same pendulum could be normal. These models are termed NonCausal herein.

A further degree of complexity can be incorporated by considering the lagged values of noncausal variables as predictors. These kinds of models are frequently termed autoregressive with exogenous inputs (ARX)* and were previously employed in dam safety.[6, 34] Specifically, the response at time $t_i$ is estimated based on the readings at $t_{i-1}$ and $t_{i-2}$, both for the variable to predict and other response variables.

One of the objectives of this work is to test the ability of all three models to detect various types of abnormalities and draw conclusions for practical purposes.

### 2.4 Case study

La Baells dam is a double-curvature arch dam located in the Llobregat river, in the Barcelona region (Spain). The crest length is 403 m, whereas the maximum height above foundation is 102 m. Monitoring data were provided by the Catalan Water Agency for the period 1980–2008. These data correspond both to environmental and response variables. In this work, the air temperature (Figure 3) and the reservoir level (Figure 4) time series were considered as inputs to a FEM. The results of this model in terms of radial displacements at the location of the pendulums were extracted and compared to the actual measurements (Figure 5). The objective was to check that the FEM could provide realistic data to generate reference time series of dam behaviour. These artificial data

### Table 1 Predictor variables considered for the causal BRT model

| Code   | Group            | Type        | Period (days) |
|--------|------------------|-------------|---------------|
| Lev007 | Hydrostatic load | Original    | 7             |
| Lev014 |                  |             | 14            |
| Lev030 |                  | Moving average | 30            |
| Lev060 | Hydrostatic load |             | 60            |
| Lev090 |                  |             | 90            |
| Lev180 |                  |             | 180           |
| Tair   |                  |             | 1             |
| Tair007|                  |             | 7             |
| Tair014|                  |             | 14            |
| Tair030| Air temperature  | Moving average | 30            |
| Tair060|                  |             | 60            |
| Tair090|                  |             | 90            |
| Tair180|                  |             | 180           |
| Rain   |                  |             | 1             |
| Rain030| Rainfall         | Accumulated | 30            |
| Rain060|                  |             | 60            |
| Rain090|                  |             | 90            |
| Rain180|                  |             | 180           |
| NDay   | Time             | Original    | -             |
| Year   |                  |             | -             |
| Month  | Season           | Original    | -             |
| n010   | Hydrostatic load | Rate of variation | 10            |
| n020   |                  |             | 20            |
| n030   |                  |             | 30            |

*The ARX model is also non-causal, in the sense that variables with non-causal relation with the outputs are included as predictors. The acronym ARX was employed to distinguish both models when necessary, although they are occasionally jointly referred to as “non-causal models”. For the sake of clarity, the capitalised version (“Non-Causal”) is used to specifically refer to the second model, excluding the ARX.
are free from any temporal variation (the reference numerical model does not vary with time; only environmental loads do).

The dam was considered as a three-dimensional solid discretised in hexahedral serendipity 27-node elements. A portion of the foundation was also included, resulting in 13,029 nodes and 2,530 elements. The thermal and mechanical problems were solved separately on the resulting finite element mesh (Figure 6), generated with the software GiD.[35] The material properties are shown in Table 2.

For the thermal problem, a transient computation was run over the 1980–2008 period with time step of 30 days. The temperature was imposed in both dam faces, with different values for the wet and dry areas. For the boundaries below the reservoir level, the temperature was considered as equal to that of the water, which in turn was estimated by means of the Bofang formula.[36] Although it allows accounting for the temperature variation with depth, a unique value was considered in this work for all the wetted boundaries, equal to that obtained for 50% depth. For the dry faces, the 30-day moving average of air temperature was imposed, to take into account the thermal inertia. The result was increased by 2° to account for the solar radiation, following the approach proposed by Perez and Martinez for Spanish dams in the northeast region.[37] The temperature evolution for the first year was repeated 4 times to ensure that the result was not influenced by the initial conditions.

### TABLE 2  Material properties considered in the FE model

| Property                        | Dam          | Foundation   |
|---------------------------------|--------------|--------------|
| Young modulus $(N \cdot m^{-2})$ | $4.76 \times 10^{10}$ | $3.10 \times 10^{10}$ |
| Poisson ratio                   | 0.25         | 0.25         |
| Density $(kg \cdot m^{-3})$     | 2,400        | 3,000        |
| Thermal conductivity $(W \cdot K^{-1} \cdot m^{-1})$ | 2.4          | 2.2          |
| Thermal expansion coefficient   | $10^{-5}$    | $10^{-5}$    |
| Specific heat $(J \cdot kg^{-1} \cdot ^\circ K^{-1})$ | 982          | 950          |

*Note. FE = finite element.*
The mechanical response was assumed to be elastic and instantaneous (without inertia); hence for each time step, the hydrostatic load correspondent to the actual reservoir level was applied.

The results of both models (thermal and mechanical) were added, and the displacement evolution at the location of the monitoring devices was extracted. The model results, which are generated in global axes, were later transformed to the local axes correspondent to the radial displacements, as measured by the monitoring devices.

Finally, weekly values were obtained via interpolation, according to the average reading frequency for the available data.

In addition to radial displacements, also the temperature evolution in the dam body was compared to observed data from several thermometers embedded in the dam body.

The goodness of fit of the FEM was computed in terms of the mean absolute error (MAE):

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

where $N$ is the number of observations, $y_i$ are the observed values, and $Fem(x_i)$ the FEM model results.

### Table 3

Discrepancy between the normal displacements, as computed with the FEM model, and those imposed in Scenario 3 for $a = 2$ mm. Mean absolute error (MAE; mm)

| Device | MAE (mm) | Device | MAE (mm) |
|--------|---------|--------|---------|
| P1DR1  | 0.61    | P6IR1  | 0.02    |
| P1DR4  | 0.52    | P6IR3  | 0.01    |
| P2IR1  | 0.10    | P5DR3  | 1.05    |
| P2IR4  | 0.13    | P5DR1  | 1.42    |

Note: FEM = finite element model.

FEM model representing a hypothetical sliding of the left abutment. For that purpose, the boundary condition at that region was set to $a$ mm both in $x$ and $y$ axes (instead of null displacement, as for the reference case).

It is important to note that the anomaly of Scenario 3 affects differently to each of the devices analysed. Because a displacement in the left abutment was imposed, the results in the left half of the dam body are anomalous. However, those in the right half are not affected. This can be observed in Figure 7, which depicts the displacement field in the dam body generated by the imposed anomaly with $a = 2$ mm.

Table 3 contains the mean absolute deviation between the reference and the anomalous time series for each device for $a = 2$ mm. Because the anomaly in Scenario 3 does not affect to some devices, those values considered as abnormal by the system will be false positives.

For each scenario, the performance of the three models considered (causal, NonCausal, and autoregressive) was analysed. 4,000 anomalous cases were generated, where the following parameters were randomly selected:

- Initial date of abnormal period
- Anomaly scenario
- Output variable
- Magnitude: 0.5, 1.0, or 2.0 mm·year$^{-1}$ for Scenario 1; 0.5, 1.0, or 2.0 mm for Scenario 2; and 1.0 or 2.0 mm for Scenario 3.

---

**Figure 7** Displacement field resulting from the anomaly in Scenario 3. View from downstream.

**Table 3** Discrepancy between the normal displacements, as computed with the FEM model, and those imposed in Scenario 3 for $a = 2$ mm. Mean absolute error (MAE; mm)
Each anomalous case was presented to all three models to compare their ability for anomaly detection. This was computed in terms of the *detection time* ($t_{\text{det}}$), defined as the elapsed time from the start of the anomaly until the first observation considered anomalous by each model, measured in days (Figure 8). Since the abnormal period was limited to 1 year, the models that did not detect any anomaly were assigned a $t_{\text{det}}$ value of 365 days.

Moreover, the effectiveness of an anomaly detection system also depends on the number of false positives (observations considered abnormal by the model, which are actually normal) and false negatives (abnormal values not detected as such by the model). The two most commonly used metrics to account for these are Precision (Equation 2) and Recall (Equation 3). In this paper, the comparison was mainly based on the $F_2$ Index (Equation 4),[14] which jointly considers precision and recall, giving more importance to the latter.

\[
\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}},
\]

\[
\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}},
\]

\[
F_2 = \left(1 + 2^2\right) \frac{\text{precision} \cdot \text{recall}}{4 \cdot \text{precision} + \text{recall}}.
\]

However, these indexes are not useful for model performance assessment when analysing the unaffected devices in Scenario 3. In these cases, there are not true positives (all records are normal, because these devices are not affected by the anomaly). Hence, both precision and recall equal zero. Nonetheless, it is highly relevant to know whether the proposed models correctly identify these records within the prediction interval. For that purpose, Scenario 3 was analysed by means of the amount of false positives, whose computation depends on the device. For those in the left half of the dam body (as viewed from upstream), which are actually anomalous, the observations above the upper limit of the prediction interval are considered as false positives, because they would imply a deviation towards upstream (although the actual anomaly corresponds to a displacement in the downstream direction). By contrast, for the unaffected devices, every record outside the prediction interval is a false positive, both above the upper limit and below the lower limit of the interval.

### 2.6 Load combination verification

In general, model accuracy is dependent on the values of the input variables. The more input data available for similar situations to that to be predicted, the more accuracy is to be expected. In dam behaviour, it will depend on the thermal and hydrostatic loads.

This effect is more important when input values are out of the training data range.[38] In particular, the accuracy of data-based models as BRTs may decrease dramatically when extrapolating.

Cheng et al.[15] defined a possible abnormal state of the dam (State 3) that “may be caused by extreme environmental values variables.” In this work, this issue was explicitly verified, and out-of-range (OOR) instances were considered as potential false positives.

This verification was carried out following an original procedure, specifically designed for the dam behaviour problem, where there are three main loads: thermal, mechanical (hydrostatic head), and temporal.

If the behaviour of the dam does not change over time, the importance of time variable is negligible. This was checked when fitting BRT models to the reference data, which correspond to time-independent dam behaviour. The inclusion of these variables is useful for retrospective analysis, as confirmed by previous studies.[11] In practice, a previously
trained model is employed to predict future values. Hence, it is obvious that the model prediction is an extrapolation in time axis and thus does not need to be verified.

As for the other two loads (thermal and hydrostatic), the simplest approach would be to check whether their values for the test period are greater (lower) than the maximum (minimum) within the training data set. However, that would not consider that both effects are coupled: The water temperature is different to that of the air, hence, the water surface elevation affects the boundary condition in the upstream dam face and, as a result, conditions the thermal response of the dam.[39]

Moreover, there is not a widely accepted agreement on what extrapolation is and how to handle it.[38] In dam behaviour modelling, it seems obvious that a hydrostatic load above the maximum in the training set is OOR. However, a more detailed definition seems appropriate to account for the “empty space phenomenon,”[40] that is, the existence of areas without training samples within the range of the inputs.

To account for this issue, the criterion employed in this work is based on the combination of both loads:

1. The training data are plotted in the (reservoir level and air temperature) plane.
2. A two-dimensional density function is computed by means of the kernel density estimation method.
3. The training instance with lower density value is localised, and the corresponding isoline is plotted.
4. The input values for the new data are plotted on the same plane. Those falling outside the isoline are considered as OOR.

With this procedure, it is taken into account that the predictive accuracy can be poor for a load combination not previously presented, even though their values, if considered separately, are within the training range. An example of this issue is presented in Figure 8.

3 | RESULTS AND DISCUSSION

3.1 | FEM accuracy

Figure 9 shows the comparison between the observed radial displacements for P1DR1 and those obtained with the FEM for the period 1994–2008. Results for other outputs are similar (Table 4). The FEM model accuracy is comparable to that obtained in previous studies with data-based models.[8]

As regards the temperature, Figure 10 shows the numerical results and the observed data for four thermometers and the January 2007 to June 2008 period. Both the devices and the time period correspond to the results published by Santillán et al.[41] who employed a highly detailed thermal model, also for La Baells dam.

Because this study does not specifically focus on predicting the thermal response, relevant simplifications were employed to generate the reference data (neglecting the variation in water temperature with depth, using a relatively large time step). Nonetheless, the temperature within the dam body was well captured.

### Table 4

| Output | MAE (mm) | Output | MAE (mm) |
|--------|----------|--------|----------|
| P1DR1  | 0.70     | P5DR1  | 0.81     |
| P1DR4  | 0.65     | P5DR3  | 1.01     |
| P2IR1  | 1.08     | P6IR1  | 0.96     |
| P2IR4  | 0.98     | P6IR3  | 0.58     |

*Note*: MEA = mean absolute error.
This, together with the results for displacements, confirm that the resulting data series mostly reproduce the dam response to the main loads. Therefore, they are representative of the normal behaviour of the dam and useful to evaluate the ability of the methodology to detect anomalies.

### 3.2 Prediction accuracy

The performance of all models on the reference data (without anomalies) was first assessed. The objectives are (a) verify the evolution of the prediction accuracy over time (as more data is included in the training set), (b) check the effect of averaging the standard deviation, (c) compare all models in terms of false positives, and (d) evaluate the efficiency of the criterion to detect OOR data.

For that purpose, the iterative process described in subsection 2.2 was followed, that is, each model was refitted yearly over an increasing training set, and the prediction interval was updated as a function of the actualised value of the weighted average of the residual standard deviation. Because the dam–foundation behaviour is time independent for the reference case, the variation in model accuracy is due to the increase of training data.

Figure 11 shows the evolution of both the raw and the weighted average of the residual standard deviation for all devices and models. Some conclusions can be drawn:
Table 5 contains the amount of false positives for all targets and models, as well as those correspondent to OOR inputs. Although the prediction interval for the causal model is wider (due to the higher residual standard deviation), it also generates a greater quantity of false positives. However, the average amount is low in all cases, as compared to the total amount of records (1,464). Moreover, the procedure to identify OOR inputs reduces the false positives by 27% for the causal model and by 45% for both the NonCausal and the ARX. As a result, the mean percentage of false positives is 8.0%, 2.8%, and 2.6%, respectively. It should be noticed that the results for the NonCausal and ARX models are lower than the theoretical percentage of values outside
the interval within 2 times the standard deviation in a normal distribution (5%).

3.3 Anomaly detection

Figure 12(a) shows the $F_2$ results as a function of the model and the anomaly magnitude $a$ for Scenarios 1 and 2. As expected, the larger anomalies were more easily detected in all cases. As for the input variables, NonCausal model performed better on average, especially for small anomalies and as compared to the causal model. Again, the inclusion of lagged variables generated a minor effect, in this case towards slightly poorer performance.

The results for Scenario 3 are more interesting to analyse, because they correspond to a realistic anomaly affecting the overall dam behaviour. Because the effect of this anomaly is different to each output, the results are presented in terms of the true detection time $t_d$ per device, that is, the elapsed time until the first record identified as a deviation towards downstream. Figure 13 shows the results.

A perfect model would feature null detection time for the affected devices (P1DR1, P1DR4, P5DR1, and P5DR3) and 365 days for the remaining (P2IR1, P2IR4, P6IR1, and P6IR3). Both the NonCausal and the ARX models showed almost perfect performance. As regards the causal model, the anomaly in the most affected devices (P5DR1 and P5DR3) is detected almost instantly but is less effective for P1DR1 and P1DR4, whose deviation from the reference behaviour is low (see Table 4). The detection time for P1DR1 and P1DR4 is around 2 months, with high variation up to 300 days. A complete assessment of the model performance requires analysing the amount of false positives. They correspond to any value outside the prediction interval for the targets in the right half of the dam body and to anomalies correspondent to deviations towards upstream for those in the left region. Figure 14 shows these results.

It can be observed that the causal model is clearly more effective in this regard: Both the NonCausal and the ARX models classify around half of the observations for the unaffected devices as abnormal (there are 52 observations in the period of analysis). This result is due to the nature of the inputs for each model. For example, the NonCausal model generates a prediction for P6IR1 based on the value of P5DR1 (among other inputs, but this is particularly important for being symmetrical within the dam body). In Scenario 3, P5DR1 deviates towards downstream with respect to the reference (training) period. Because that input is anomalous, the resulting prediction is also wrong. In this case, the model interprets that the value of P6IR1 falls in the upstream side of the prediction interval.

This issue is highly relevant, because the final aim of the system is not only to detect a potentially anomalous behaviour but also to support the correct identification of the cause and then the decision making. In fact, similar results would have been obtained had the devices been analysed jointly in Scenarios 1 and 2: A real deviation towards downstream in some

![Figure 13](image-url) Detection time (days) per target and model for Scenario 3. ARX = autoregressive with exogenous inputs
device is (in general) correctly identified by the noncausal models, but that same value would generate an incorrect prediction for other devices, of opposite sign.

Causal models do not give these spurious results, because they predict the dam response only based on the external variables, at the cost of a generally higher detection time.

A straightforward option to avoid this behaviour is to discard noncausal models. However, their good performance for detecting true anomalies suggests that they can be useful overall.

As an alternative, the outputs whose value is identified as anomalous by a noncausal model can be removed from the input set. The model requires retraining, but it can still offer accurate results, thanks to the flexibility of BRTs.

A new set of 240 cases was run for Scenario 3 and the NonCausal model. The results shown in Figure 15 confirm that the removal of abnormal variables is effective against false positives, while maintaining the ability for anomaly detection. The model performance is only poorer for P2IR1 (unaffected by the anomaly in Scenario 3): The detection time is lower than 365 days, which indicates the existence of false positives.

FIGURE 14 False positives per target and model for Scenario 3. ARX = autoregressive with exogenous inputs

FIGURE 15 Detection time and false positives per target for Scenario 3 and the NonCausal model, once the anomalous variables are removed from the input set
Nonetheless, the average detection time is still 270 days, and the total amount of false positives is lower than 10%.

This approach was implemented in a new visualisation tool, which was developed to present the results for all devices involved. It is based on the Shiny library[42] and includes two plots for each model (Figure 16). First, each device is plotted on its actual location within the dam body, with a symbol that is a function of the deviation between prediction and observation for the date under consideration. Then, the evolution of observations and predictions for the most recent period is plotted for one device selected by the user. Figure 16 shows the application interface for one of the anomalies from Scenario 3. It can be observed that the anomaly is correctly localised.

With this tool, the user jointly receives the overall information on all devices under consideration, and a more detailed plot of the selected output, where the value of the deviation, as well as the trend, can be observed. In this version, devices whose residuals are lower than 2 times the standard deviation are plotted in green; those between 2 and 3 times are depicted in yellow, and those above 3 times are shown in red. The shapes correspond to the direction of the deviation (upstream or downstream), as interpreted by each model. This criterion can be tailored to the user preferences.

4 | SUMMARY AND CONCLUSIONS
A methodology for early detection of anomalies in dam behaviour was presented, which includes a prediction model based on BRT, a criterion for detecting anomalies based on the residual density function, and a procedure for realistic estimation of the prediction interval. Also, extraordinary loads are identified by jointly considering the two most important external loads (hydrostatic load and temperature).

Causal models (which only consider external variables) and noncausal (including both internal and lagged variables as predictors) were compared in terms of detection time for three different anomaly scenarios. The results showed that noncausal models are more effective for the detection of anomalies, both affecting to isolated devices (Scenarios 1 and 2) and those resulting from an overall malfunction of the dam (Scenario 3).

In the case study considered, the inclusion of lagged variables had minor effect both in the model accuracy and the detection time. This suggests that the NonCausal model (without lagged variables) might be a better choice due to its higher simplicity.

Causal models were more robust as regards the precision (when accounting for false positives). In abnormal periods, the prediction of noncausal models for unaffected devices is
often wrong because it is partially based on anomalous data (that from the devices actually affected by the anomaly). This type of behaviour is a consequence of the nature of the model itself and is the price to pay in exchange for a greater ability for early detection of anomalies.

However, an updated version of the NonCausal model, where the anomalous variables are removed from the input set, avoided the above-mentioned issue and showed to be as effective for anomaly detection as the NonCausal and even more robust against false positives than the causal model. Hence, this approach is the best option to provide useful information to the dam safety managers. To that end, it was implemented in an interactive online tool, which shows the devices whose behaviour is interpreted as potentially abnormal by the predictive model, together with the plot of the evolution of predictions and observations for all relevant outputs.

This tool can be used as a support for decision making, because it facilitates the identification of a potential deviation from normal behaviour. Thus, it can be used as an indicator to generate a warning that might lead to intensify the dam safety monitoring activity. Nonetheless, all relevant decisions influencing dam safety should be made by an expert and capable engineer, based on the analysis of all the relevant information available.

ACKNOWLEDGEMENTS

The authors thank Carlos Barbero, dam safety manager at the Catalan Water Agency, for providing the monitoring data.

The research was 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] International Commission on Large Dams, Dam surveillance guide, Technical Report B-158, ICOLD 2012.
[2] J. Mata, A. Tavares de Castro, J. Sá da Costa. Struct. Control Health Monit. 2014, 21(3), 423.
[3] J. Mata. Eng. Struct. 2011, 3(3), 903.
[4] V. Ranković, N. Grujović, D. Divac, N. Milivojević. Struct. Saf. 2014, 48, 33.
[5] H. Su, Z. Chen, Z. Wen. Struct. Control Health Monit. 2016, 23(2), 252.
[6] V. Ranković, N. Grujović, D. Divac, N. Milivojević, A. Novaković. Eng. Struct. 2013, 35, 107.
[7] F. Salazar, R. Morán, M. Á. Toledo, E. Oñate. Arch. Comput. Meth. Eng. 2015, 24, 1.
[8] F. Salazar, M. A. Toledo, E. Oñate, R. Morán. Struct. Saf. 2015, 56, 9.
[9] B. K. Myers, D. H. Scofield, Providing improved dam safety monitoring using existing staff resources: Fern Ridge Dam case study, in Proc. 28th Annual USSD Conf., Portland, Oregon, USA 2008.
[10] H. Yu, Z. Wu, T. Bao, L. Zhang. Sci. China Technol. Sc. 2010, 53(4), 1088.
[11] F. Salazar, M. Á. Toledo, E. Oñate, B. Suárez. Eng. Struct. 2016, 119, 230.
[12] J. Mata, N. S. Leitão, A. T. de Castro, J. S. da Costa. Comput. Struct. 2014, 142, 45.
[13] V. J. Hodge, J. Austin. Artif. Intell. Rev. 2004, 22(2), 85.
[14] I.-S. Jung, M. Berges, J. H. Garrett, B. Poczos. Adv. Eng. Inform. 2015, 29(4), 902.
[15] L. Cheng, D. Zheng. Adv. Eng. Softw. 2013, 57, 48.
[16] S. Gamse, M. Oberguggenberger. Struct. Control Health Monit. 2017, 24, e1859.
[17] C.-Y. Kao, C.-H. Loh. Struct. Control Health Monit. 2013, 20(3), 282.
[18] F. Li, Z. Wang, G. Liu. Struct. Saf. 2013, 43, 12.
[19] C.-H. Loh, C.-H. Chen, T. -Y. Hsu. Struct. Health Monit. 2011, 10(6), 587.
[20] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth & Brooks, Monterey, CA 1984.
[21] J. H. Friedman. Ann. Stat. 2001, 9, 1189.
[22] G. Ridgeway, Generalized boosted models: A guide to the gbm package, http://CRAN.R-project.org/package=gbm, R package vignette, 2007 (accessed: March 16, 2017).
[23] J. R. Leathwick, J. Elith, M. P. Francis, T. Hastie, P. Taylor. Mar. Ecol. Prog. Ser. 2006, 321, 267.
[24] J. Elith, J. R. Leathwick, T. Hastie. J. Anim. Ecol. 2008, 77(4), 802.
[25] R. E. Schapire, in Nonlinear Estimation and Classification, Springer, New York 2003, 149–171.
[26] A. Michelis. 2012, Traditional versus non-traditional boosting algorithms, Master’s Thesis, University of Manchester, Manchester.
[27] R Core Team, R: A language and environment for statistical computing, http://www.R-project.org/, R Foundation for Statistical Computing, Vienna, Austria, 2013.
[28] G. Ridgeway, gbm: Generalized boosted regression models, R package version 2.1. 2013.
[29] F. Salazar, J. M. González, M. A. Toledo, E. Oñate, in Proc. 8th European Workshop Struct. Health Monit., Bilbao, Spain 2016.
[30] R. J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, 2013, OTexts: Melbourne, Australia. Available at: http://otexts.org/ffp/ (accessed March 16, 2017).
[31] S. Arlot, A. Celisse, Stat. Surv. 2010, 4, 40.
[32] M. Herrera, L. Torgo, J. Izquierdo, R. Pérez-García. J. Hydrol. 2010, 387(1), 141.
[33] J. H. Friedman, J. J. Meulman. Stat. Med. 2003, 22(9), 1365.
[34] P. Palumbo, L. Piroddi, S. Lancini, F. Lozza, in Proc. Sixth ICOLD Benchmark Workshop on Numer. Anal. Dams, Salzburg, Austria 2001.
[35] R. Ribó, M. A. R. Pasenau, E. Escolano, J. S. P. Ronda, L. F. González, GiD reference manual, in CIMNE, Barcelona 1998.
[36] Z. Bofang. Dam Eng. 1997, 8, 13.
[37] J. L. Pérez, E. Martínez. Rev Obras Públicas 1995, 3349, 79. [in Spanish].
[38] T. Ebert, J. Belz, O. Nelles, in 2014 IEEE Symp. Comput. Intell. Dat. Min. (CIDM), IEEE, Orlando, Florida, USA 2014, 310–314.
[39] M. Tatin, M. Briffaut, F. Dufour, A. Simon, J. -P. Fabre. Eng. Struct. 2015, 91, 26.
[40] M. Verleysen, in Limitations and Future Trends in Neural Computation (Ed: S. Ablameyko), IOS Press, Amsterdam 2003, 141–162.

[41] D. Santillán, E. Salete, D. J. Vicente, M. Á. Toledo. J. Eng. Mech. 2014, 140(11).

[42] W. Chang, J. Cheng, J. J. Allaire, Y. Xie, J. McPherson, shiny: Web application framework for R, http://CRAN.R-project.org/package=shiny, R package version 0.13.2, 2016 (accessed: March 16, 2017).

How to cite this article: Salazar F, Toledo MA, González JM, Oñate E. Early detection of anomalies in dam performance: A methodology based on boosted regression trees. Struct Control Health Monit. 2017;24:e2012. https://doi.org/10.1002/stc.2012