Promoting Explainability in Data-Driven Models for Anomaly Detection: A Step Toward Diagnosis

Quentin Dollon[1], Paul Labbé[2], and François Léonard[3]

[1,2,3] *Hydro-Quebec Research Institute (IREQ), Varennes, Quebec, Canada J3X 1S1*
Dollon.quentin2@hydroquebec.com
Labbe.paul@hydroquebec.com
Leonard.francois@hydroquebec.com

ABSTRACT

This study introduces a data-driven model for anomaly detection in hydroelectric generating units. After an initial course of training, a monitoring stream is deployed that compares asset behaviour to the expected behaviour. Training and monitoring coexist for some time, allowing early monitoring of the asset. Efforts were made to extract as much statistical explainability as possible during development of the model. This renders the approach more reliable and consistent for decision-making support and helps to reduce false positive alerts. Examples of how this tool can be used in industry to make a step toward asset diagnosis are given.

1. INTRODUCTION

Anomaly detection has become a critical task in industry and serves various purposes, including reliability analysis, safety assurance and asset health monitoring. Data-driven models are often used for anomaly detection given their ability to learn patterns from data and identify behaviours that deviate from the learned patterns (Sutharssan, Stoyanov, Bailey, & Yin, 2015; Tsui, Chen, Zhou, Hai, & Wang, 2015). They are also simple to implement since they do not rely on complex physical models to make predictions. A major limitation of these models, however, is their lack of explainability, which hinders the diagnosis of detected anomalies.

Explainability provides transparency and interpretability, allowing stakeholders to understand the reasons for detected deviations from normal behaviour. In the absence of explainability, it is challenging to determine why a particular realization was classified as abnormal. Without an understanding of the underlying reason for an anomaly, it is difficult to make a reliable diagnosis, which can result in missed opportunities for preventing or mitigating damage caused by the anomaly. Explainability can also help in detecting false positives and false negatives, especially in distinguishing between abnormal behaviours and sensor failures or unseen operating regimes.

Hydro-Quebec is Canada's largest power utility and a major player in the global hydropower industry. Hydro-Québec generates more than 99% of its electricity from hydroelectric generating units. Power grid sustainability thus depends heavily on effective health monitoring of these assets. This paper introduces a data-driven semi-supervised algorithm for anomaly detection with emphasis on statistical explainability. This explainability differs from that of traditional explainable models, which build on physics to interpret observations. Here, the goal is to track sources of deviation through statistics to explain why the software believes that an anomaly has occurred. This semi-supervised model is not a diagnostic tool, however, because its sole output is insufficient for determining the root causes of a problem. It does nonetheless offer a bridge toward such tools by providing clues about the origin of failures. In addition, the proposed model is able to start monitoring after a very short initial training using a limited dataset. As more data is incorporated in the algorithm, confidence increases and so does sensitivity.

In the following section, data preparation and pipeline construction (including relevant feature extraction, curation, and scaling) are described. Next, the data-driven approach used to model asset behaviour is presented (Léonard, Merleau, Tapsoba, & Gagnon, 2019), focusing on its adaptivity, that is, its ability to evolve as data is fed to the algorithm. The detection metric is then introduced as a multidimensional statistical deviation called Hypersphere Realization Deviation (HRD). HRD can be seen as a measure of the multidimensional distance between a realization and model predictions. The expected distance is not zero: a probability shell develops around predictions within which most normal observations lie. This is due to the presence of noise in the data. Lastly, the explainability features of the algorithm are high-

lighted and some practical examples are given to demonstrate the algorithm's versatility and performance.

2. DATA PREPARATION

An in house Extract-Transform-Load (ETL) pipeline was constructed to prepare the data, which consists mainly of asynchronous time series extracted from OSIsoft's PI system, SCADA and other databases. The series are deemed asynchronous because the time delay between successive measurements is positive but random. Certain features of interest, such mean values, RMS, peak-to-peak and spectral components, are pre-computed during data acquisition and are readily extracted when available. The pipeline can then compute additional user defined features of interest from extracted data before creating a synchronised data frame from selected time series. The pipeline can then construct additional columns by applying user defined operators on existing columns before user defined filters are applied to remove rows from the data frame.

In particular, the pipeline needs to filter out transients and dead times since the algorithm is trained on steady states because these states establish the normal behaviour of the asset. Transient states are identified by measuring the deviation to time averaged values prior to the synchronisation step. Transients and dead times are then simply filtered out in the last step.

The rows of the synchronised data frame can be seen as a series of snapshots, each representing the state of the asset at a given time. At time m, snapshot $\boldsymbol{z}^m$ contains two types of information: the independent variables $\boldsymbol{x}^m \in \mathbb{R}^I$ that form the operating condition domain and the independent variables $\boldsymbol{y}^m \in \mathbb{R}^D$ inducing the asset response manifold. The data is scaled using a Min-Max scaler. For a new realization, this transform is updated as:

$$\begin{cases} z_{j\,\max}^m = \max\left(z_{j\,\max}^{m-1}, z_j^m\right) \\ z_{j\,\min}^m = \min\left(z_{j\,\min}^{m-1}, z_j^m\right) \end{cases}, \forall j \in [\![1, I+D]\!] \quad (1)$$

3. CLUSTER-BASED KRIGING

The model used to predict asset behaviour takes a two-stage approach. A clustering algorithm is deployed to parcel out the operating condition domain dynamically. Once the data are reduced, kriging is used to interpolate between clusters and predict expected behaviour at a specific position in the operating domain.

3.1. Stream Clustering

Clustering is an unsupervised machine-learning technique used to organize a cloud of points into a limited number of collections, called clusters. A cluster represents asset be-

haviour in the vicinity of a given operating condition. Stream clustering is a variant used in monitoring that is able to process data continuously without needing the entire dataset before the domain is partitioned (Zubaroglu & Atalay, 2021).

Because we want to group data with similar operating regimes, asset response is ignored during clustering and only independent variables are provided to the algorithm. Clustering provides a set of clusters $\{\mathcal{C}_l, l \in [\![1, L]\!]\}$ characterized by population $|\mathcal{C}_l|$, centroid coordinates $\boldsymbol{x}_l, \boldsymbol{y}_l \in \mathbb{R}^I \times \mathbb{R}^D$ and by associated deviations $\boldsymbol{\sigma}_{x,l}, \boldsymbol{\sigma}_{y,l} \in \mathbb{R}^I \times \mathbb{R}^D$ (assuming uncorrelated dimensions). Since we relied solely on first and second statistical moments, we implicitly model the empirical distribution with Gaussian families. This is justified by maximum entropy theory, giving Gaussian family of probability as the Shannon information maximizer (Jaynes, 1978). To avoid indefinite creation of clusters and ensure sufficient statistical information (population) in each cluster, a limit $L_{\max}$ was imposed on the number of clusters. In the authors' experience, 30 to 70 clusters are generally sufficient to obtain accurate modelling. As $L \ll M$ (M being the length of the time history), clustering allows a significant reduction of computational burden.

The preliminary stage of training consists in seeding the model. Seeding is a two-step process. During inflation, the α_{init} first realizations are allocated to a unit cluster. A deflation step is then applied during which α_{merged} clusters are merged together. At the end of the seeding, the operating condition domain is partitioned into $\alpha_{\text{init}} - \alpha_{\text{merged}}$ regions.

After initialization, cruise training starts. During this phase, realizations are incorporated into the model using the workflow depicted in Figure 1. At each iteration, the characteristic average square radius of the clusters needs to be computed using equation (2). This represents the average dispersion of clusters in the operating domain, and is used during data assimilation.

$$r^2 = \frac{1}{\sum_{l=1}^L |\mathcal{C}_l| - 1} \sum_{l=1}^L \left(|\mathcal{C}_l| - 1\right) \boldsymbol{\sigma}_{x,l}^T \boldsymbol{\sigma}_{x,l} \quad (2)$$

For any inbound realization $\boldsymbol{z}^m$, the following operations are allowed on clusters (Leonard, 2011):

- Merge a realization to a cluster using Welford's algorithm (Welford, 1962). Merging is completed when the squared distance of the realization to its closest cluster, $d_{m,\min}^2$, is less than $\alpha_{\text{incl}} r^2$, where α_{incl} is a truncation factor (generally set to 2) excluding unlikely candidates.
- Reject a realization from the training circuit. It is crucial for clustering to not learn abnormal behaviours. For this reason, any realization that violates equation (16) is rejected. A realization assigned to a saturated cluster is rejected as well (next bullet).
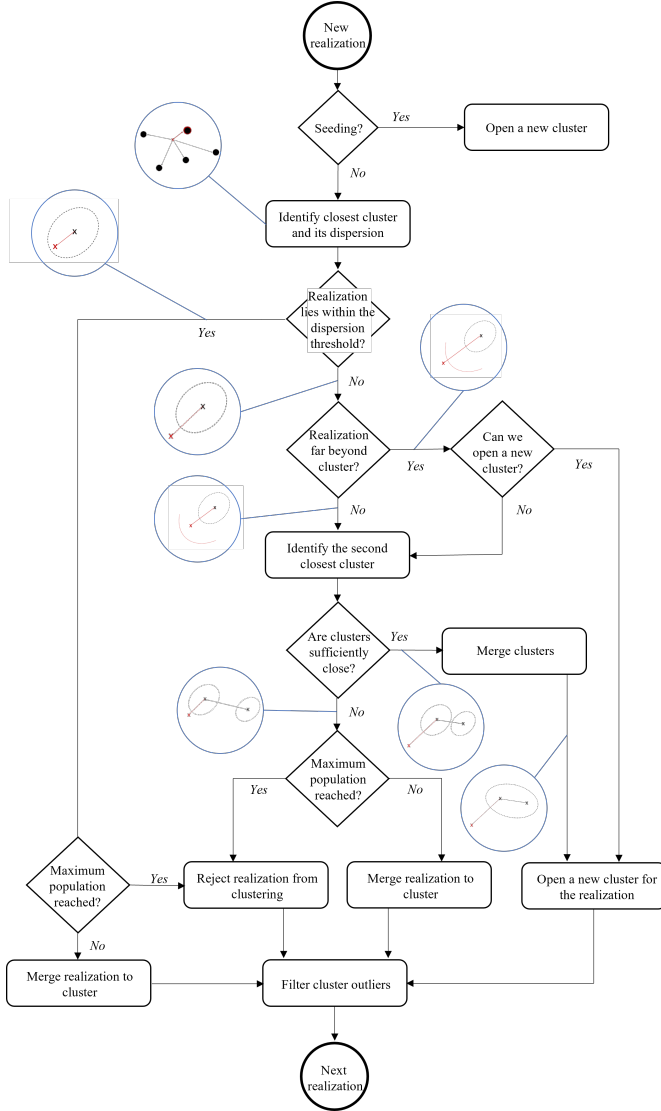
Figure 1. Clustering workflow.

- Saturate a cluster, meaning that we stop training one densely populated cluster. A maximum population is allowed because assets get progressively damaged during operation, resulting in slight but continuous deviation from healthy behaviour. To detect such slow changes, they must not be learnt, and cluster saturation prevents assimilation of such slow drifts over time.

- Open a new cluster. When $L < L_{\max}$ and the realization is far from any other cluster, $d_{m,l}^2 > \alpha_{\text{new}} r^2$ for all $l \in [\![1, L]\!]$, a new cluster is opened for these unseen operating conditions.

- Merge two clusters using Parallel algorithm (Chen, Golub, & Leveque, 1979). When a new realization needs a cluster to be opened but $L = L_{\max}$, the algorithm tries to merge two adjacent clusters if their distance is less than $\alpha_{\text{incl}} r^2$.

- Discard cluster outliers. Any cluster with a very small population that is not used during a given number of iterations is automatically discarded.

In sum, centroids and dispersions in the independent domain are used to decide whether a realization should be incorporated in the cluster or not. Later, independent and dependent centroids are used to predict "normal" asset behaviour when kriging. The dispersion in the dependent domain represents the behaviour reproducibility and is used in calculating the detection threshold.

3.2. Dual Kriging

Kriging models are used to interpolate between clusters to obtain predictions in each dependent dimension at the current operating point. Kriging is known to be the best linear unbiased predictor (BLUP) (Smith, 2001). Universal kriging is a variant used to model weak stationarities with deterministic trends. The trend is modelled as a linear combination of operating conditions using a monomial basis. Let $\boldsymbol{y}_j \in \mathbb{R}^L$ be the L cluster centroid positions in the jth dependent dimension. The kriging model is formulated as

$$\boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{\eta}_j + \boldsymbol{\delta}_j, \quad \boldsymbol{X} = \begin{pmatrix} 1 & \boldsymbol{x}_1^T \\ \vdots & \vdots \\ 1 & \boldsymbol{x}_L^T \end{pmatrix} \in \mathbb{R}^{L \times I+1} \quad (3)$$

where $\boldsymbol{\eta}_j$ is a random variable used to capture spatially correlated aleatory effects, and $\boldsymbol{\delta}_j$ is independent exogenous white noise that is known as nugget and used to smooth interpolations. Coefficients $\boldsymbol{\beta}_j \in \mathbb{R}^{I+1}$ are used to model the deterministic drift. At an unseen operating condition $\boldsymbol{x}^m$, the kriging model expresses as $y_j(\boldsymbol{x}^m) = (1 \ \boldsymbol{x}^{m\,T})\boldsymbol{\beta}_j + \eta_j^m + \delta_j^m$. The covariance matrices of the random variables are given by

$$\mathcal{C}\begin{pmatrix} \boldsymbol{\eta}_j \\ \eta_j^m \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma} & \boldsymbol{\gamma} \\ \boldsymbol{\gamma}^T & \sigma_0^2 \end{pmatrix}, \quad \mathcal{C}\begin{pmatrix} \boldsymbol{\delta}_j \\ \delta_j^m \end{pmatrix} = \begin{pmatrix} \boldsymbol{G}_j & \boldsymbol{0} \\ \boldsymbol{0} & g_{0j}^2 \end{pmatrix} \quad (4)$$

Quantities $\boldsymbol{\Gamma}, \gamma, \sigma_0^2 \in \mathbb{R}^{L \times L} \times \mathbb{R}^L \times \mathbb{R}^+$ are obtained using the so-called semi-variogram and implicitly depend on certain parameters $\boldsymbol{\theta}$. The semi-variogram models the evolution of a dependent variable in the independent domain. The experimental variogram must be fitted with an analytical "authorized" model used to calculate covariance matrices. Nugget covariances $\boldsymbol{G}_j, g_{0j}^2 \in \mathbb{R}^{L \times L} \times \mathbb{R}^+$ have diagonal structures and represent the reproducibility error at a given location. They are different for each dependent variable. For prediction, kriging assumes a linear structure of the following form:

$$\hat{y}_j(\boldsymbol{x}^m) = \boldsymbol{\lambda}_{jm}^T \boldsymbol{y}_j \tag{5}$$

To prevent propagation of deterministic weights $\boldsymbol{\beta}_j$ in the sequel, the following constraint is imposed:

$$\boldsymbol{\lambda}_{jm}^T \boldsymbol{X} = (1 \ \boldsymbol{x}^{m\,T}) \tag{6}$$

Hence, prediction error writes as $\epsilon_j^m = \eta_j^m + \delta_j^m - \boldsymbol{\lambda}_{jm}^T(\boldsymbol{\eta}_j + \boldsymbol{\delta}_j)$, from which prediction variance (7) is derived. Predictor $\hat{y}_j(\boldsymbol{x}^m)$ is found by minimizing prediction variance under constraint (6):

$$\sigma_{jm}^2 = \sigma_0^2 + g_{0j}^2 - \boldsymbol{\lambda}_{jm}^T\boldsymbol{\gamma} + \boldsymbol{\lambda}_{jm}^T(\boldsymbol{\Gamma} + \boldsymbol{G}_j)\boldsymbol{\lambda}_{jm} \tag{7}$$

Constrained minimization is performed by introducing Lagrange multipliers $\boldsymbol{\nu}_{jm} \in \mathbb{R}^{I+1}$. Eventually, $\boldsymbol{\lambda}_{jm}$ satisfies

$$\begin{pmatrix} \boldsymbol{\Gamma} + \boldsymbol{G}_j & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_{jm} \\ \boldsymbol{\nu}_{jm} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\gamma} \\ \begin{pmatrix} 1 \\ \boldsymbol{x}^m \end{pmatrix} \end{pmatrix} \tag{8}$$

The right-hand term in equation (8) depends on current operating conditions; the system must then be solved for any new operating conditions. Dual kriging is a computational efficient variant that reparametrizes the problem in a spatially independent way (Journel & C.J., 1979). It is a global interpolator where all clusters are used regardless of their distance from the regression point. The simplest way to obtain the dual representation of kriging is to submit equation (8) in (5). The dual regression is obtained as:

$$\hat{y}_j(\boldsymbol{x}^m) = \begin{pmatrix} \boldsymbol{\gamma}^T & (1 & \boldsymbol{x}^{m\,T}) \end{pmatrix} \begin{pmatrix} \boldsymbol{\psi}_j \\ \boldsymbol{\phi}_j \end{pmatrix} \tag{9}$$

where the dual coefficients $\boldsymbol{\psi}_j$ and $\boldsymbol{\phi}_j$ are obtained from the dual kriging system,

$$\begin{pmatrix} \boldsymbol{\Gamma} + \boldsymbol{G}_j & \boldsymbol{X} \\ \boldsymbol{X}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\psi}_j \\ \boldsymbol{\phi}_j \end{pmatrix} = \begin{pmatrix} \boldsymbol{y}_j \\ \boldsymbol{0} \end{pmatrix} \tag{10}$$

The main advantage of dual reformulation is that dual coefficients $\boldsymbol{\psi}_j, \boldsymbol{\phi}_j$ can be calculated once and for all and then used for any operating condition. In primal kriging regression complexity is $O\left(L^2(I+2)^2\right)$, while in dual kriging it is reduced to $O\left(L\right)$. This allows almost instantaneous estimation of normal behaviours. However, obtaining the kriging error with low computational burden can be tedious (discussed latter in section 4.3.2).
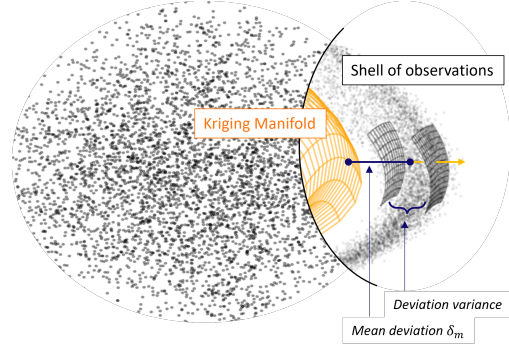

Figure 2. Shell of realizations around the kriging prediction.

4. HYPERSPHERE REALIZATION DEVIATION

Predictions need to be compared to actual observations through an appropriate metric. We use a *Hypersphere Realization Deviation* metric (HRD). This statistical measure synthesizes the multidimensional residual into a single scalar value. The concept of shell of observations, which arose naturally when deriving the metric, is introduced first below. Next, the HRD is described as well as an adaptive method for determining a responsive detection threshold above which an anomaly is suspected.

4.1. Shell of observations

The shell of observations is closely connected to the notion of Euclidean distance between a random variable and its expected value. Consider a multidimensional stochastic process with zero mean and a given positive semi-definite covariance matrix. The expected value of the Euclidean distance of realizations is then non-zero[1]. For instance, it is well known that the Euclidean distance of an uncorrelated n-dimensional Gaussian random vector follows a chi distribution with n degrees of freedom. The expected value of such a distribution is $\sqrt{2}\Gamma((n+1)/2)/\Gamma(n/2)$, which is strictly positive for any $n \geq 1$.

The concept of shell of observations states that the Euclidean distance between a realization $\boldsymbol{y}_j^m$ and its kriging prediction $\hat{\boldsymbol{y}}_j(\boldsymbol{x}^m)$ always deviates by a characteristic length δ, as shown in Figure 2. It is possible to evaluate δ experimentally. The residual vector between realizations and predictions is defined as

$$\Delta\hat{\boldsymbol{y}}(\boldsymbol{x}^m) = \begin{pmatrix} \hat{y}_1(\boldsymbol{x}^m) - y_1^m \\ \vdots \\ \hat{y}_D(\boldsymbol{x}^m) - y_D^m \end{pmatrix} \tag{11}$$

---

[1]Since a distance is always positive, the only way the expected distance could be null is if all realization distances were null, meaning the distribution is a delta function. This is inconsistent with the existence of a dispersion imposed by the covariance matrix.

The residual Euclidean distance is obtained as

$$d_m = \sqrt{\Delta\hat{\boldsymbol{y}}^T(\boldsymbol{x}^m)\Delta\hat{\boldsymbol{y}}(\boldsymbol{x}^m)} \qquad (12)$$

By averaging these deviations in equation (13), quantity $\hat{\hat{\delta}}_m$ is obtained as an estimate of δ, the expected distance between realizations and predictions. This quantity can be seen as a measure of the noise corrupting the data. Set $\mathcal{S}_m$ is the sampling subset and contains indices of the realizations used to compute statistics. It is obtained from equation (16) and used to prevent learning of abnormal behaviours.

$$\hat{\hat{\delta}}_m = \frac{1}{|\mathcal{S}_m|}\sum_{i\in\mathcal{S}_m} d_i \qquad (13)$$

4.2. HRD Indicator

The idea behind the HRD is to have a way of comparing deviations from expected distance δ (Leonard, 2021). For the mth realization, the HRD metric ρ_m is defined as

$$\rho_m = d_m - \hat{\hat{\delta}}_{m-1} \qquad (14)$$

In concise terms, the HRD analyses the variation of the realization deviation to the expected distance. It is a geometrical interpretation of the multidimensional information. HRD follows a centered distribution whose variance gives the thickness of the shell of observations. This variance will be calculated in the next section.

As defined in equation (14), HRD gives equal weight to information carried by each dimension of the dependent manifold. However, this can result in overrepresentation of false positives due to sensor failure. This is due to the common source of certain features used as dependent variables. In fact, if one extracts mean value, RMS, peak-to-peak and the three most significant spectral components (frequency, amplitude, phase) from one sensor, there are then 12 dependent variables fed by the same device. This means the weight associated with failure of the sensor will be 12 times higher than it should be. To correct this bias, a redundancy factor was introduced in the analysis to balance the weight of each channel so that total information weight for each sensor used for feature extraction is 1.

4.3. Responsive Detection Threshold

The detection threshold ρ_{lim} in equation (15) represents the upper limit for the HRD before an alert is sounded. It is updated with each new iteration to adapt to the confidence of the current prediction, increasing when HRD uncertainty grows and decreasing when it diminishes. Constructing this cutoff requires proper quantification of uncertainties throughout the model. There are two types of uncertainties: asset response
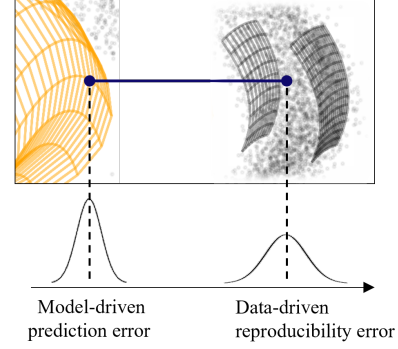

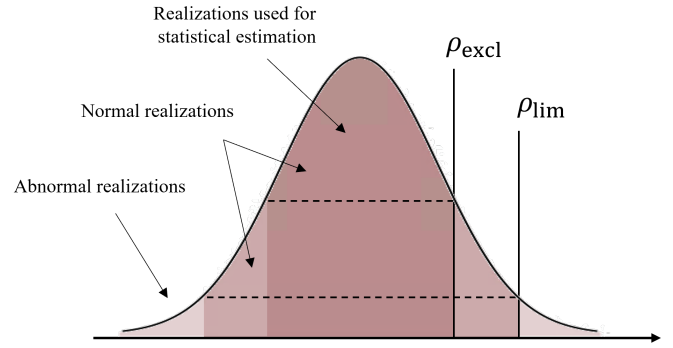Figure 3. Prediction and observation variances in the model.

Figure 4. HRD bounds used in the model.

reproducibility σ_{rep}^2 and kriging error σ_{ck}^2, as shown in Figure 3. For the mth realization

$$\rho_{\text{lim}}(m) = \sqrt{\alpha^2\sigma_{\text{rep}}^2(m) + \sigma_{\text{ck}}^2(m)} \qquad (15)$$

where α determines the reproducibility confidence interval and is generally set to 4. A more exclusive threshold is used to determine which realizations should be used in the statistical estimates, *i.e.*, to build subset $\mathcal{S}_m$:

$$\rho_{\text{excl}}(m) = \beta\sigma_{\text{rep}}(m) \qquad (16)$$

Usually, $\beta = 2$. Relations between these different bounds is illustrated in Figure 4.

4.3.1. Reproducibility error

Variance σ_{rep}^2 is determined experimentally:

$$\hat{\sigma}_{\text{rep}}^2 = \frac{1}{|\mathcal{S}_m| - 1}\sum_{i\in\mathcal{S}_m}\rho_i \qquad (17)$$

This is an underestimate, however, because rejection of abnormal realizations is equivalent to truncating the distribution. Underestimating the variance results in lowering ρ_{lim}

and ρ_{excl}, which in turn leads to smaller variances and so on. The estimated variance needs to be corrected using a factor $\gamma(\beta) > 1$

$$\hat{\sigma}^2_{\text{rep}}(m) = \frac{\gamma(\beta)}{|\mathcal{S}_m| - 1} \sum_{i \in \mathcal{S}_m} \rho_i \tag{18}$$

At greater dimensions, the HRD distribution tends to Gaussian. With Gaussian distributions, this correction is given by

$$\gamma(\beta) = \frac{\text{erf}\left(\beta/\sqrt{2}\right)}{\text{erf}\left(\beta/\sqrt{2}\right) - 2\beta\varphi(\beta)} \tag{19}$$

where φ is the standard Gaussian probability density function and erf is the error function. With $\beta = 2$, one has $\gamma(2) \approx 1.14$. The uncorrected variance is underestimated by 14%.

4.3.2. Model error

Reproducibility error accounts for uncertainties due to interpolation and model error in previous realizations. However, it does not address current interpolation conditions. For instance, if interpolation is done near a widespread cluster, or worse far away from any cluster, uncertainty will be considerable. Theoretically, this uncertainty is encapsulated in the kriging variance given in equation (7). However, dual kriging does not provide this value, and variance must be recovered differently. In this section, some conservative rules are proposed to obtain a reasonable estimation of $\sigma^2_{\text{ck}}(m)$ in equation (15). Two terms must be distinguished in the model variance:

$$\sigma^2_{\text{ck}}(m) = \sigma^2_{\text{c}}(m) + \sigma^2_{\text{k}}(m) \tag{20}$$

The term $\sigma^2_{\text{k}}(m)$ is due to interpolation between clusters, while $\sigma^2_{\text{c}}(m)$ is associated with spatial discretization of the clustering. The formulae shown below result from extensive empirical studies and years of trial and error. These developments proved that the following formulae yield acceptable measures of sensitivity with respect to clustering:

$$\sigma^2_{\text{c}}(m) = r^2 \varepsilon^2_{\text{c}}(m; \kappa) \tag{21}$$

where r^2 is defined in equation (2) and

$$\varepsilon^2_{\text{c}}(m; \kappa) = \frac{\sum_{l=1}^{L} |\mathcal{C}_l| - 1}{\sum_{l=1}^{L} (|\mathcal{C}_l| - 1) \, w_l^2(m; \kappa)} \tag{22}$$

The term $w_l(m; \kappa)$ gives a weight for each cluster to the total variance:

$$w_l(m; \kappa) = \frac{r}{\max\left(d_{m,l} - \kappa r; 10^{-5}\right)} \tag{23}$$

This uncertainty is solely dependant on information about the operating condition domain. When the distance between a non-unitary cluster and the realization exceeds κr, then $w_l(m; \kappa)$ starts decreasing and $\sigma^2_{\text{c}}(m)$ increases. Conversely, when this distance drops below κr for one cluster, the related uncertainty becomes negligible. A floor of 10^{-5} is imposed to not singularize the error. For sensitivity with respect to the kriging process, the following measure performs well:

$$\sigma^2_{\text{k}}(m) = r^2 \sum_{j=1}^{D} \varepsilon^2_{\text{k}j}(m) \tag{24}$$

where $\varepsilon^2_{\text{k}j}(m)$ is the contribution of the jth dependent variable, expressed as

$$\varepsilon^2_{\text{k}j}(m) = \frac{2}{\boldsymbol{\gamma}^T \mathbb{1}} \frac{\left(\boldsymbol{\gamma}^T \boldsymbol{G}_j \boldsymbol{\psi}_j\right)^2}{\boldsymbol{\gamma}^T \left(\boldsymbol{G}_j \boldsymbol{\psi}_j\right)^2} \tag{25}$$

5. EXPLAINABILITY FEATURES OF THE ALGORITHM

5.1. Algorithmic reversibility

Algorithmic reversibility refers to the ability to trace data used in predictions back to their source. The proposed metric compares each inbound realization to the model, which was trained on a data history. When analyzing global deviation of asset behaviour, the user naturally wonders which periods of the data history were used to build predictions. For instance, are the data involved recent or do they date back to earlier years during the same season? Do the historical segments used concentrate on specific dates or are they spread out over time? Are these segments numerous or very limited?

In the data reduction step, clusters accumulate data from different time periods. Each cluster then has a temporal distribution of the incorporated realizations. This distribution is discrete to save memory. The temporal distribution associated with a prediction corresponds to the sum of all the temporal distributions over clusters weighted by their kriging influence (weights ϕ_j in the interpolation). When interpolating near a cluster, there is almost no influence from other clusters and the temporal distribution corresponds to that of the cluster. On the other hand, when studying new operating conditions, the time history is based on the contribution of neighbouring clusters.

5.2. Adaptive detection threshold

As explained in section 4.3, the detection threshold is not static but is set according to model confidence. When the model is confident, *i.e.*, makes predictions near a cluster (well-known operating region) or in high reproducibility regions, then the confidence interval becomes very narrow and the metric very sensitive. Conversely, when the model makes

predictions in unknown operating conditions or in regions with weak reproducibility, the confidence interval increases to not capture false positives. This behaviour is easy to observe in practical cases like those described in the following section.

This adaptive threshold is crucial for early monitoring. Early monitoring refers to the period when monitoring starts while training is still ongoing. We do this because operating conditions strongly depend on the season (water levels, temperatures), and passing through all operating conditions takes at least a year, which is too long. At the beginning of monitoring, clusters are sparse and realizations often show new operating conditions, making model confidence low. But as clusters get denser and the operating domain better known, confidence increases and the HRD becomes more sensitive (see Figures 6 and 8).

5.3. Feature importance analysis

Feature importance analysis aims at determining a score for each element of the realization. More precisely, it makes it possible to quantify how much each feature contributes to an observed deviation. This performance analysis is easy to incorporate in the approach we propose, as HRD construction involves a preliminary multidimensional distribution that compares the current realization to expected values. Statistical distance of the realization from the prediction in each dimension is computed. This quantity, called the z-score, is obtained as a Mahalanobis distance, giving a measure of the distance in number of standard deviations. The z-score is then scaled to give relative contributions to HRD. When the algorithm raises an anomaly, the participating features are determined as those above the z-score threshold, and their degree of contribution is used for diagnosis. The z-score threshold used to characterize a feature as out-of-distribution is generally fixed at 3.

Feature importance analysis is undoubtedly the most important explainability feature of the model and the one most used in practice. It can be used to detect sensor failure (not found by redundancy analysis) if a deviation is entirely triggered by features of one sensor or to identify the set of incriminated sensors and spatially locate the anomaly. It can also be used, when relevant, to determine the operating conditions under which an anomaly occurs.

6. EXAMPLES FROM INDUSTRY

The model is currently being deployed on the Hydro-Quebec hydro generating unit fleet. In this section, examples are given of HRD use with actual cases encountered in recent years. The assets studied are hydro generating units located in Québec, Canada. A schematic diagram of such a unit with its main components is illustrated in Figure 5.
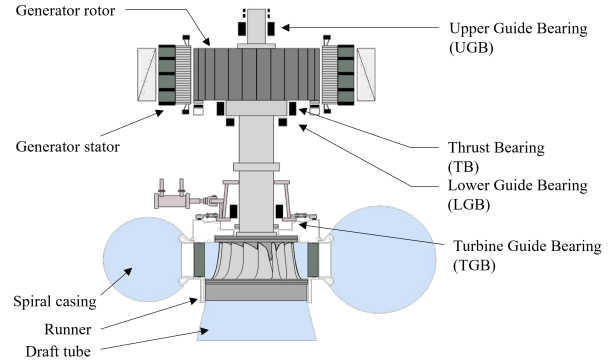

Figure 5. Schematic of a hydroelectric generating unit.

We generally use four features to represent the operating conditions of a generating unit: upstream and downstream water levels, guide vane opening and ambient temperature. The dependent features are extracted from row sensor measurements, including two orthogonal displacements at each guide bearing (UGB, LGB and TGB), stator temperature, oil and babbit temperature for the bearing cooling systems, thrust bearing acceleration and output power. Generally, the dependent features extracted are RMS (A), spectral RMS (B), synchronous response (C), second harmonic (D), peak-to-peak (E) and mean value (F) or instantaneous value (G).

6.1. Case 1: When everything goes well

The first hydro generating unit studied was a low-head propeller turbine that generally outputs 120 MW to the grid. The low head of around $25m$ is compensated by high water inflow. The unit has two guide bearings but no upper guide bearing. This unit is healthy and faces no particular problems when operating.

The metric was first used to study the behaviour of this non-problematic unit. Results are shown in Figure 6, where three regions are distinguished. No HRD was calculated for the initiation region from January to the end of February 2019 because there were not enough realizations to compute reliable statistics. This period was followed by the early monitoring phase, during which clusters are sparse. Given the data scarcity, the model is not confident and the detection threshold was raised accordingly. With time, clusters become populated, asset response becomes better known and the detection threshold drops accordingly. During steady-state operation, detection levels remain frozen for most monitoring. This is because units generally operate at similar operating points and these regions are well represented by the model.

Sometimes, however, grid stability requires the unit to operate under exotic conditions. As shown in 6, these unseen regions are reflected by a loss of confidence and a threshold peak. Weak overshoots of the HRD appeared in June 2020 and August 2021. We focus here on the June 2020 event. When the
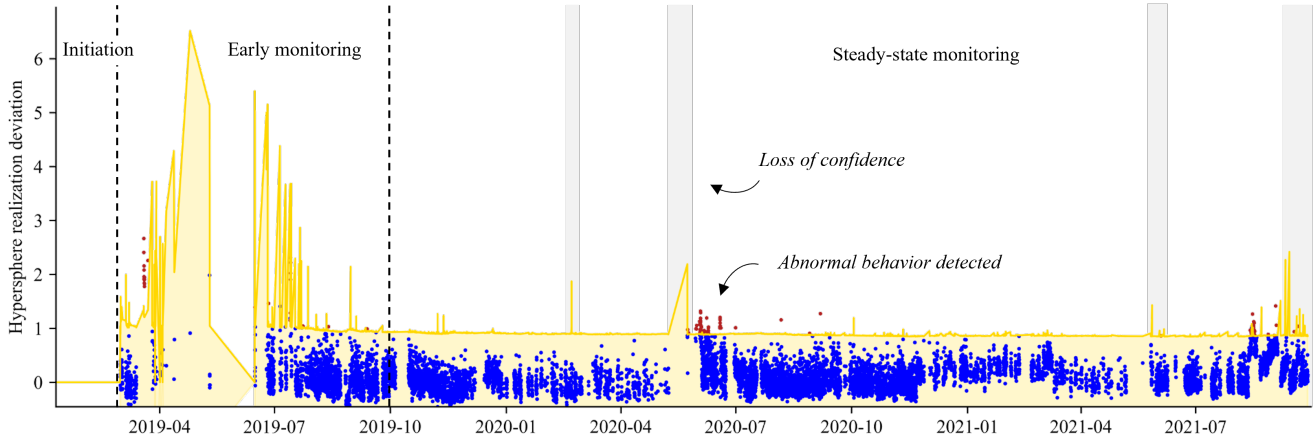
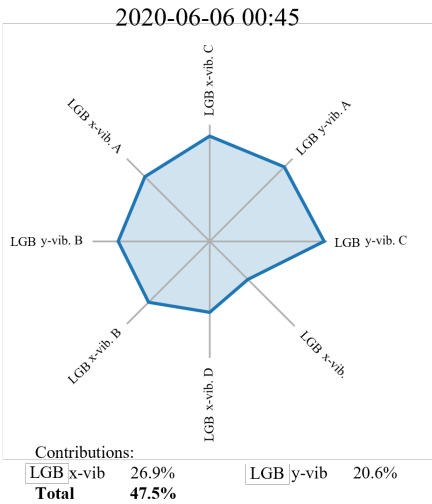Figure 6. Example of HRD metric behaviour when an asset works normally.

Figure 7. Radar plot of out-of-distribution features for detected deviation.

metric exceeds the threshold, it means a statistically significant number of features are out-of-distribution. These features and their relative contribution can be plotted, as shown in Figure 7. This radar plot brings explainability and reveals that the anomaly in this case was due to a small increase in lower guide bearing (LGB) vibrations. The x-direction and y-direction sensors at the LGB contributed to the deviation by 26.9% and 20.6% respectively. Such corroboration from two closely correlated sensors gives confidence in the measured vibration level. As there is no UGB, the information from the LGB suggests a minor unbalance at the stator. This stealth anomaly disappeared in June after routine maintenance when, among other things, the stator was cleaned.

6.2. Case 2: Early failure detection

In July 2018, a failure alarm triggered by the protection system resulted in emergency shutdown of one of our hydro generating units after only a few years of operation. Inspection revealed major damage to the runner blade orientation system, with fractured stoppers and pivots and a cracked housing structure. This was mainly due to an inappropriate runner design. The repairs requiredimmobilization of the asset for nearly two years, which meant nearly two years of generation loss as well. The HRD was used to conduct an a posteriori analysis of the power plant's data history, the aim to evaluate its anomaly detection performance. Indeed, earlier detection of the machine malfunction would have meant more cost-effective repairs, a shorter downtime and the possibility of planning maintenance during energy demand gaps.

Figure 8 shows the HRD metric calculated for a dataset spanning January 2016 to July 2018. Figure 9 shows a radar plot of out-of-distribution features on specific dates, numbered from (1) to (4) in Figure 8. The most significant damage occurred in mid-July, when the HRD was 15 times the detection threshold. However, the first signs of abnormal behaviour date back to August 2017 (see (1) in Figures 8 and 9), one year before the emergency shutdown. The anomaly was mainly indicated by disproportionately high acceleration at the thrust bearing, which explained 40% of the observed deviation. Synchronous responses at the TGB and the LGB were also out-of-distribution, but their contribution to the HRD was marginal (around 5%).

Another period of anomaly began at the end of October 2017 and lasted for over five months. During this time, a large number of features were out-of-distribution: 20 features are abnormal on the radar plot in Figure 9 (see (2)), contributing to 87% of the total deviation. The abnormalities were mainly related to measured TGB and LGB displacements, but output
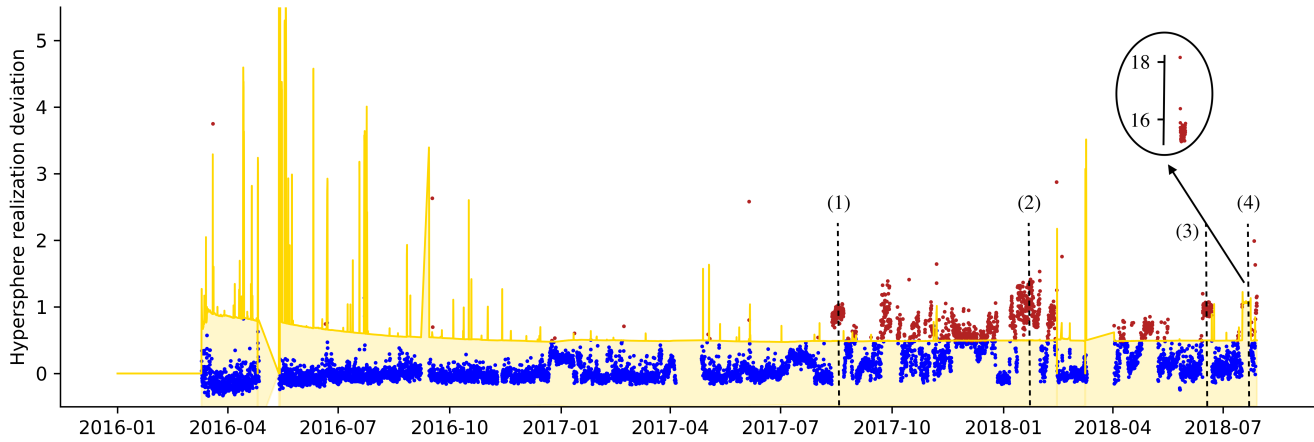
8

Figure 8. Example of HRD metric behaviour when an asset is suffering from a worsening failure: emergency unit shutdown triggered by ultimate safety protocol protection system in August 2017.
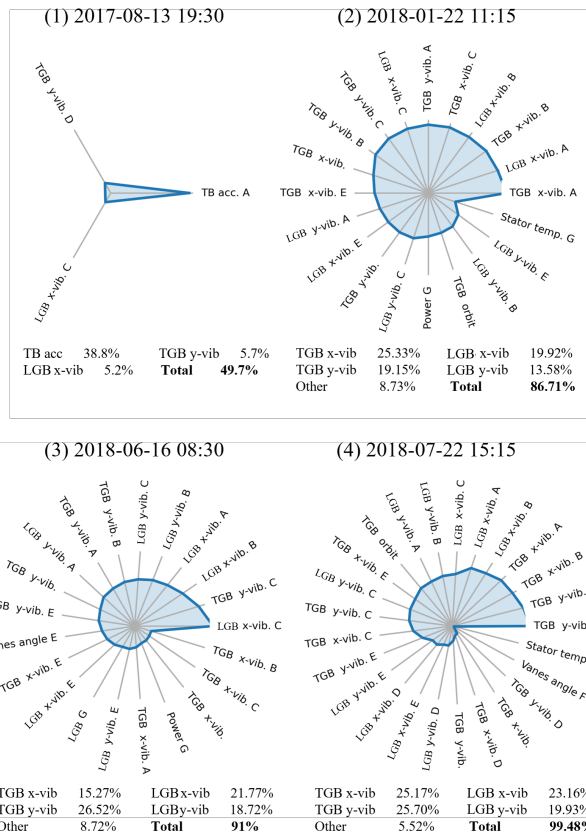

Figure 9. Radar plot of out-of-distribution features for dates when failure progressed.

power was also unusual. UGB vibrations, on the other hand, were normal and did not raise any issues. The multisensorial analysis showed that the lower part of the shaft line was vibrating abnormally. The global deviation gathers contributions from six sensors, so the possibility of a sensor failure

could be discounted. The HRD rose above detection threshold once again in April 2018, when blade position was found to be out-of-distribution in addition to the earlier contributing features, suggesting even more strongly that there was a runner blade issue. Finally, in July 2018, the loads on the structure became intolerable and the machine broke down. It is interesting to note that as the damage propagated, the out-of-distribution features became the only sources of deviation, meaning they moved farther and farther away from their expected distributions.

Our conclusion from this synthetic analysis is that our monitoring model was able to detect the first signs of failure a full year before this major accident. Four periods of anomaly ranging from one week to five months over the course of the year preceding the failure indicated a gradual degradation of performance. Was there enough information to predict the exact failure and the remaining useful life? Of course not. However, the evidence provided by the model was more than sufficient to instigate a visual inspection of the runner that would have undoubtedly led to discovery of damage.

6.3. Case 3: Dealing with unexplored regimes

To illustrate how the model deals with unexplored regimes, a unit that had to operate with an abnormally low upstream water level during winter 2023 (see Figure 10) was selected. Upstream water level was relatively constant with an average value of 31.75m from the start of training in October 2020 until January 2023, when it dropped rapidly to 31m and remained there until March 2023. This difference might seem insignificant to a neophyte, but change in head (potential energy of water flow) has a major impact on fluid configuration and affects not only the mechanical behaviour of the runner but also the power produced.

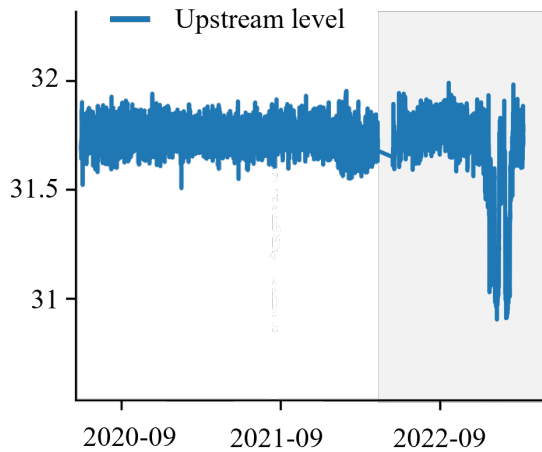When an asset starts operating in new regimes, the metric

9

Figure 10. Upstream level defining one independent variable during training: level dropped by 1m during winter 2023.
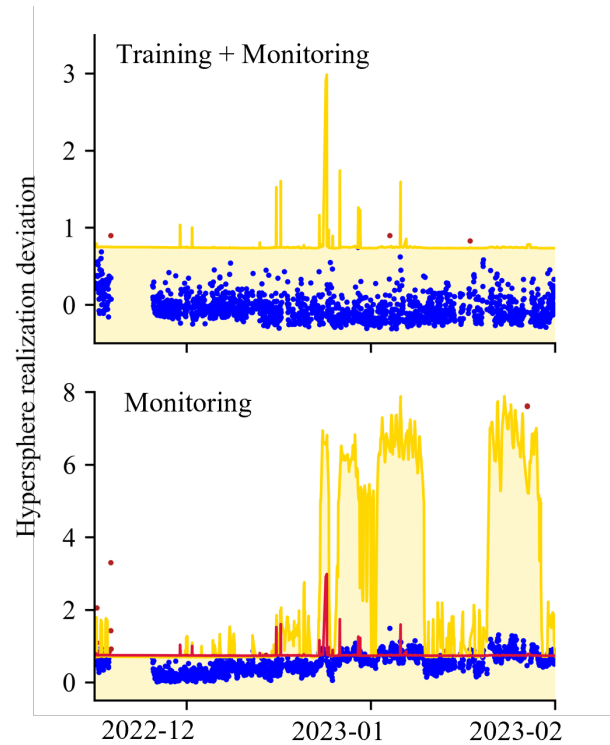

Figure 11. Metric behaviour when dealing with new operating regimes. Upper figure: training and monitoring coexist and new regime becomes new cluster, restoring confidence. Lower figure: training is stopped and model loses confidence. Red line represents detection threshold without consideration of model error.

can have two distinct behaviours. If the model is still being trained, an unseen operating condition raises a stealth loss of confidence but the new regime is rapidly clusterized in the model, which rapidly recovers confidence. This is shown in the upper part of Figure 11. However, when the model stops being trained, the unseen operating conditions remain unknown to the model and the confidence interval grows dramatically, as shown in the lower part of Figure 11.

When investigating asset behaviour under operating conditions far from any cluster, the model error proposed in section 4.3.2 is crucial to prevent false alarms. The red line in Figure (11) represents the reproducibility error of the realizations. If this variance alone is considered, the HRD would clearly exceed the threshold when no anomaly should be reported. When model error is taken into account, however, the HRD remains within the confidence interval.

7. CONCLUSION

This paper describes the development of a data-driven algorithm for asset health monitoring with emphasis on explainability. Explainability strengthens software reliability and provides a bridge to diagnosis. It also enables automatic removal of false positives due to loss of confidence or sensor failures, an essential aspect of any failure detection metric as failures are rare in industrial systems. It would be detrimental to introduce a monitoring system that triggers more false alarms than real failure alarms. Thanks to the proposed metric, an early monitoring of assets is permitted, albeit with lower sensitivity.

As the examples show, the HRD metric is currently being used to monitor rotor dynamics of generating units. There is little electrical or hydraulic input to the realizations. Most

features come from displacement sensors located on bearing housings. Including a broader spectrum of feature types is an ongoing development which could lead to a richer asset health condition panel.

REFERENCES

Chen, T., Golub, G., & Leveque, R. (1979). Updating formulae and a pairwise algorithm for computing sample variances. *Technical Report STAN-CS-79-773, Department of Computer Science, Stanford University*.

Jaynes, E. (1978). Where do we stand on maximum entropy? In *Proceedings of the maximum entropy formalism conference.* (MIT, Boston, United States)

Journel, A., & C.J., H. (1979). *Mining geostatistics*. New York Academic Press.

Leonard, F. (2011). *Dynamic clustering of transient signals.* United States Patents. (US 2014/0100821)

Leonard, F. (2021). *Quantitative analysis of signal related measurements for trending and pattern recognition.* United States Patents. (US 10,902,088 B2)

Léonard, F., Merleau, J., Tapsoba, D., & Gagnon, M. (2019, May). Hydro-turbine monitoring: from self-learned

equipment behavior to a single global deviation indicator. In *22nd iris rotating machine conference.* (New-Orleans, United States)

Smith, R. L. (2001). *Environmental statistics*. University North Carolina.

Sutharssan, T., Stoyanov, S., Bailey, C., & Yin, C. (2015). Prognostic and health management for engineering systems: a review of the data-driven approach and algorithms. *Journal of Engineering*, *2015*(7), 215-222.

Tsui, K., Chen, N., Zhou, Q., Hai, Y., & Wang, W. (2015). Prognostics and health management: a review on data driven approaches. *Mathematical Problems in Engineering*, *2015*.

Welford, B. P. (1962). Note on a method for calculating corrected sums of squares and products. *Technometrics*, *4*, 419-420.

Zubaroglu, A., & Atalay, V. (2021). Data stream clustering: a review. *Artificial Intelligence Review*, *54*, 1201-1236.

BIOGRAPHIES

Quentin Dollon received his master's degree in 2017 from École des Mines in Albi, France. He pursued doctoral research on Bayesian statistics for inverse problems in mechanical engineering, developing less restrictive approaches for ambient modal analysis and proposing a model to better understand pressure and displacement harmonic fields in hydroelectric turbines. He was awarded a Ph.D. with highest distinction by the École de Technologies Supérieure of the Université du Québec and Institut National des Sciences Appliquées of Lyon in 2021. In the mean time, he joined Hydro-Quebec's research institute, working in the prediction and reliability unit on development of digital twins for asset management and condition monitoring.

Paul Labbé received his M.Sc.A. in mechanical engineering in 1993 from École Polytechnique de Montréal in Québec, Canada. He then joined a computational fluid dynamics (CFD) group at CERCA, a research centre specializing in high performance computing, where he worked on NURBS-based modeling of boundary represented 3D geometries, mesh generation, error estimation and adaptivity of FEM and FVM problems. He joined the CFD group at Maya HTT in 2003 and worked with the advanced meshing group for CAE in a leading commercial software. In 2008, he joined a research team working on hydraulic turbines at Hydro Quebec's research institute as a specialist in mesh generation, error estimation and adaptivity. He completed his Ph.D. in computer engineering in 2011 at École Polytechnique de Montréal in Québec, Canada. Since 2018, he has been working on digital twins of hydro generators and data-driven models for anomaly detection.

François Léonard received his M.Sc. in physics in 1981 from École Polytechnique de Montréal in Québec, Canada. He then joined a research team working on wind turbines at Hydro-Québec's research institute as a specialist in instrumentation and signal processing. Among other achievements, he developed a special modal tool, the "Zmodal", for estimating the low damping modes of wind turbines. From 1987 to 1989, he wrote the code for a monitoring system now used for every large hydraulic turbine at Hydro-Québec. From 1990 to 1995, he worked on hydro-turbine vibration diagnosis and kriging of databases in monitoring systems. From 1995 to 2010, he worked on vibroacoustical monitoring of transformer tap-changers and switchgear, vibroacoustical crack detection of cap and pin insulator porcelain and partial-discharge detection and location in underground power cable networks. From 2010 to 2015, he was on the team that developed a transformer bushing monitoring system using GSP synchronization for phase measurement. Since 2015, he has been working on smart meter grid data analysis and has also returned to working on hydro turbine monitoring, this time using data clustering and kriging cluster models. He wrote the patents behind many of the leading-edge products developed at Hydro-Québec.