Dam Behaviour Prediction Using an Ensemble of Bayesian Dynamic Linear Model and Bayesian LSTM Networks

Bhargob Deka*, Van-Dai Vuong*, James-A. Goulet
Department of Civil Engineering, Polytechnique Montreal, Montreal, Canada

Patrice Côté, Benjamin Miquel
Direction of Dams and Infrastructure Expertise, Hydro-Québec, Montreal, Canada

ABSTRACT: In this paper, we present our submission to the ICOLD benchmark for the two pendulum datasets (CB2 & CB3). Our approach relies on the ensembling of a *Bayesian dynamic linear model* (BDLM) along with Bayesian *long-short-term memory* (LSTM) neural networks that use the *tractable approximate Gaussian inference* method (TAGI) for learning its parameters. We provide through our probabilistic ensembling method, the explainability of BDLMs as well as the accuracy and ease of use of Bayesian LSTM. Although the benchmark focusses on prediction accuracy and threshold value definition for the purpose of anomaly detection, one should keep in mind that this way of envisioning anomaly detection is only one approach among many others. We show in this paper that with our probabilistic regime switching method we expect to be able to detect anomalies of 0.5 mm for CB2 and 0.15 mm for CB3, where both cases, anomalies can develop over the span of years.

1 INTRODUCTION

Sensor-based structural health monitoring (SHM) is an established tool for informing dam owners and managers about the occurence of abnormal events as well as the general condition of the structure. Although it is a routine task to measure structural responses such as displacements, inclinations, pressure or flow rates using commercial technologies, it is much harder to extract information and knowledge from data. In the context of dam monitoring, the hydrostatic-seasonal-time (HST) method (Salazar et al., 2017) is the most common approach in order to model the dependency between structural responses and water level, seasonal components and time. The HST method has passed the test of time, nevertheless, it has several limitations; (1) it has a limited forecasting capacity when the relationship between the explanatory variables or their components are non-linear, or affected by a phase shift; (2) it is an offline method, i.e., the model is built using a training set so that it requires periodic retraining in order to adapt to new conditions not covered during training. This affects the capacity to detect anomalies that are building up over years as model re-training will capture a part of the anomaly in the model itself. The research community is still figuring out what are the options in order to overcome these limitations. In this context, the ICOLD workshop on dam behaviour prediction aims at comparing various methods with respect to their *predictive capacity*, *anomaly detection capacity* and *interpretability*.

In this paper, we present our submission to the ICOLD benchmark for the two pendulum datasets (CB2 & CB3). Our approach relies on the ensembling of a *Bayesian dynamic linear model* (BDLM) (Gaudot et al., 2019) along with Bayesian *long-short-term memory* (LSTM) neural networks (Goodfellow et al., 2016) that rely on the *tractable approximate Gaussian inference* method (TAGI) (Goulet et al., 2021) for learning its parameters. BDLMs enables non-linear dependencies between model sub-components, is an online method capable of updating itself as new

*equal contribution

data comes in, is inherently probabilistic so that it can handle epistemic and aleatory uncertainties, and it allows explaining the dependencies within the model. LSTMs excel at forecasting without requiring feature engineering regarding the interactions between structural responses, explanatory variables and other latent variables and its coupling with the TAGI method makes it inherently probabilistic as well. Ensembling (Sagi & Rokach, 2018) is a common approach in order to aggregate the predictions from several models with the objective of improving the accuracy through the cancellation of the model errors in the case they are statistically independent.

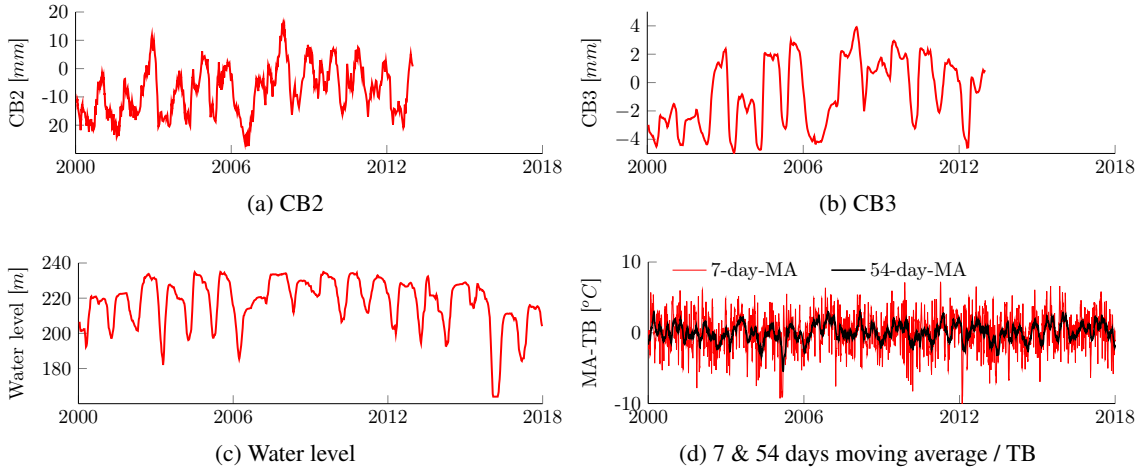
The paper is organized as follow: Section 2, presents the dataset employed as well as the preprocessing steps we applied on the data. Section 3 presents the methodologies behind the BDLM, LSTM, and ensembling methods. Section 4 presents the results regarding validation, forecasting, and model interpretation where we also investigate anomaly detectability.

2 DATASETS & PREPROCESSING

In this paper, we are building models for the pendulum time-series CB2 and CB3, measuring the dam's radial displacement [mm] from the bottom to crest, and foundation to bottom, respectively. In order to model these time series, we rely on the reservoir water level [m] as well as the temperature data TB [°C]. The data acquisition for the displacements CB2/3 has been made with an average frequency of 1.5 week, whereas the average frequency of the reservoir water level as well as the temperature TB is daily. We use daily data in our models both in training and forecasting which means that there are many missing data points in the CB2/CB3 displacement datasets.

For BDLM models, the water level data below 196m have been truncated to that value in order to account for the physical constrain associated with the bottom of the dam. In addition, instead of using the raw data itself, we account for the thermal inertia of the dam by using a $\{1,7,14\}$ (CB2) & $\{14,28,54\}$ (CB3) days moving averages for the residual of temperature TB where the yearly periodic pattern has been removed. Here, for each sensor, we selected the moving average periods which which led to significant contribution for the displacement predictions among the set $\{1,7,14,28,54\}$. Note that the one day moving average is equivalent to the raw data.

For LSTM models, we use the raw data of the reservoir water level and the temperature TB. This is because the corrected pattern introduced by the truncation of the water level is detrimental to the accuracy of the LSTMs prediction. Furthermore, LSTM models can take into account the lagging effect of the temperature on the dam's displacement automatically by using a lookback period larger than one. Figure 1 presents the data that is employed as input in order to build the BDLM and TAGI-LSTM models.



(a) CB2 (b) CB3

(c) Water level (d) 7 & 54 days moving average / TB

Figure 1. CB2/3 displacements, water level, and examples of moving averages for the temperature TB.

3 METHODOLOGY

This section presents the theoretical foundations behind Bayesian dynamic linear models, the coupling between tractable approximate Gaussian inference and LSTMs, as well as the Gaussian mixture method for aggregating the predictions from multiple models.

3.1 Bayesian Dynamic Linear Models (BDLM)

Linear regression and neural networks are categorized as parametric methods because the relationships within the model are controlled by the estimation of parameters. On the other hand, BDLMs fall in the non-parametric category as the relationships within the model are learnt probabilistically through constraints describing the transition of hidden state variables through time, as well as their observability. For example in order to model the position x_t in time t of an object in free-fall, rather than trying to adjust the parameters of a function in order to fit through observations of the tuples (time, position), i.e. a parametric approach, BDLM would model the dependency through time $h(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ using the classic kinematic equations for the hidden states $\boldsymbol{x} = [x, \dot{x}, \ddot{x}]^\intercal$; the position x, speed $\dot{x}$ and acceleration $\ddot{x}$, and their observability by defining that only the position is observable, i.e. $y_t = x_t$. From these constraints on the transition and observability, we can then employ the Kalman filter (Kalman, 1960) (i.e., the Gaussian conditional equations) in order to infer the posterior probability density function $f(\boldsymbol{x}_t|y_1, y_2, \cdots, y_t)$ of the hidden states given the data.

As stated in introduction, the main advantage of such an approach is that it allows updating the model online as the data become available, without needing to re-learn the model parameters. In practice, one can rely on a collection of predefined sub-components, each modelling a specific behaviour present in a time series, and which can be assembled together in order to create powerful, yet simple, models. Another key aspect of BDLMs is their capacity to handle regime switches over time, where models describing different regimes (e.g., a constant regime versus a linearly changing one) can compete against each other and are ranked according to their prior probability, the probability to switch from one regime to another, and the likelihood of each at explaining the data. This regime switching algorithm is the backbone of anomaly detection in the context of BDLMs (Nguyen & Goulet, 2018a; Khazaeli et al., 2021), whereas a switch between regimes can be used as a proxy indicating the presence anomalies. The complete details regarding the BDLM theory can be found in (Goulet, 2020), examples of its application to SHM datasets can be found in (Nguyen et al., 2019; Goulet & Koo, 2018; Nguyen & Goulet, 2018b; Goulet, 2017; Nguyen & Goulet, 2017), and the OpenBDLM library (Gaudot et al., 2019) can be found on GitHub (`https://github.com/CivML-PolyMtl/OpenBDLM`).

For this submission, the architecture of our model can be subdivided according to each time series, i.e., displacement, water level and temperature moving average. The selection of the model's components and their dependencies, were defined iteratively in order to remove any distinguishable pattern from the model residual term. The water level uses a local level to model the average value, a local trend in order to extract the long-term non-periodic tendency ($\approx$ 5 years), and an autoregressive process to capture the short-term ($\approx$ 1 year) non-periodic changes in water level. The temperature is modelled using a local level to model the average value, a Fourrier-form periodic component to extract the long-term stationary pattern and an a white-noise process to capture the non-periodic changes in temperature. The displacement time series CB2/3 are modelled using a local level to represent the average value, two state-based non-linear dependencies on the water level (1) mean-centered values and (2) its long-term pattern, a linear dependency over the non-periodic changes in temperature, and an autoregressive process in order to capture the time-dependent model errors. The mathematical formulation for the matrices defining the transition and observation models are presented in appendix A and the BDLM code for reproducing the results presented in this paper has been made available on GitHub (`https://github.com/CivML-PolyMtl/OpenBDLM/tree/ICOLD_benchmark`).

3.2 TAGI-Long Short-Term Memory neural networks (TAGI-LSTM)

LSTM is the classic neural network architecture for modelling time-series data. It models the dependency between explanatory variables and target outputs using a cell state enabling to consider long-term dependencies, layers of hidden variables defining the neural networks and gates (i.e., forget, input and output) enabling the combination of the information coming from the hidden and cell states. A key advantage of LSTM over BDLM or HST methods is that it does not require a specific architecture setup for defining the possible dependencies with respect to explanatory variables, thus enabling it to be quickly applied to a large number of time series.

The parameters of LSTMs are typically learnt deterministically using gradient-based optimization. The drawback of such an approach is that it disregards the epistemic uncertainty associated with parameter estimation. In order to overcome this limitation, we rely on the tractable approximate Gaussian inference method (TAGI) (Goulet et al., 2021) in order to perform Bayesian estimation for the LSTM network parameters. The specific network architecture and the hyperparameters employed in this submission are presented in appendix B.

Like other neural network architectures, LSTM networks are sensitive to the values employed to initialize model parameters before their estimation. Given the ease to evaluate multiple models having different initial model parameters, we choose to combine ten models in order to further improve the prediction accuracy. The ensembling method to combine these ten models along with the BDLM one is presented in the next subsection

3.3 Gaussian Mixture Ensembling

The ensembling method we use in this submission relies on the moment matching Gaussian mixture of models (Runnalls, 2007). Here, we want to aggregate the BDLM and ten LSTM models where each has a Gaussian output as characterized by their respective expected value μ_i and variance σ_i^2, making them natively suited for the Gaussian mixture (GM). A GM combines several Gaussian probability density functions according to the probability associated with each model. In the case of the moment matching GM, we approximate the resulting mixture distribution by a Gaussian random variable whose moments $(\mu_{\texttt{GM}}, \sigma^2_{\texttt{GM}})$ matching those of the true mixture distribution and which can be computed using the relations

$$\begin{aligned} \mu_{\texttt{GM}} &= \sum_{i=1}^{\texttt{N}} w_i \mu_i \\ \sigma^2_{\texttt{GM}} &= \sum_{i=1}^{\texttt{N}} w_i \left[\sigma_i^2 + (\mu_i - \mu_{\texttt{GM}})^2\right], \end{aligned}$$

where for $\texttt{N}$ models, the GM expected value is the weighted sum of the individual μ_i, and the GM variance is the weighted sum of the variance σ_i^2 plus a term to account for the discrepancy between each model's expected value.

In a Bayesian context, the weights should be computed according to their posterior probability $w_i = p(m_i|\mathscr{D})$ as defined by

$$p(m_i|\mathscr{D}) = \frac{p(\mathscr{D}|m_i) \cdot p(m_i)}{\sum_i p(\mathscr{D}|m_i) \cdot p(m_i)}.$$

Here, we rely on the simplifying assumption that $p(\mathscr{D}|m_i) = \ln \mathscr{L}(m_i)^{-1}$ is equal to the inverse log-likelihood of each model measured between 2012-2013, whose values are reported in appendix C. The prior probability $p(m_i)$ for the BDLM model is equal to 0.5, and to 0.05 for each of the TAGI-LSTM models making their aggregated prior probability also equal to 0.5.

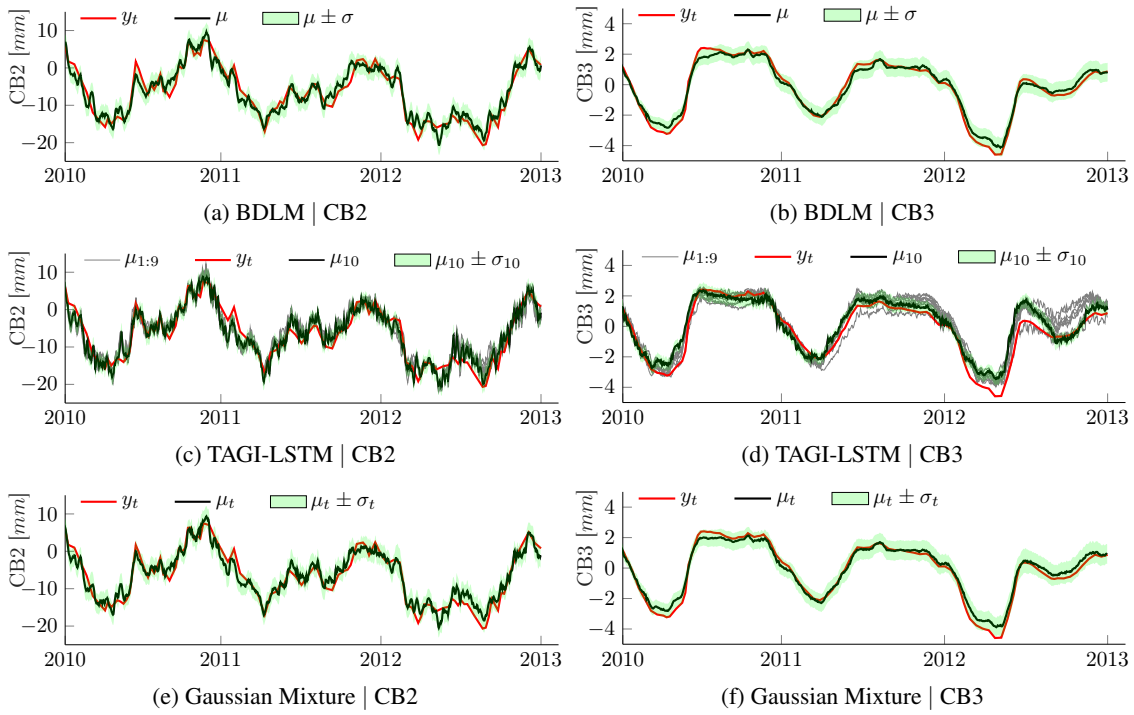
4 RESULTS

We divided the presentation of the results into three parts; first, we present the relative performance of each individual model, i.e., BDLM vs TAGI-LSTM by training each of them on a subset of the

training data available, and then predicting the last three years of data available; second, we present the forecasted data aggregating the prediction of one BDLM and 10 TAGI-LSTM models; third, we present the model interpretation where we identify the sources and nature of the dependencies between time series. Finally, we present how the regime switching capacity of BDLM is better at detecting anomalies than threshold-based alarm triggers.

4.1 Validation

Figure 2 compares the predictions for the last three years of the training data available, obtained for each models and for the Gaussian mixture of all models. These results show that both methods offer a comparable performance with a slight edge for the BDLM method. In terms of computational time, both methods are comparable with a total training time in the order of an hour. Once trained, both models can be use to predict with a computational time in the order of a minute.



(a) BDLM | CB2 (b) BDLM | CB3

(c) TAGI-LSTM | CB2 (d) TAGI-LSTM | CB3

(e) Gaussian Mixture | CB2 (f) Gaussian Mixture | CB3

Figure 2. Comparative performance of BDLM, 10 instances of TAGI-LSTM, and the Gaussian mixture from BDLM and 10 instances ($\{\mu_1, \mu_2, \cdots, \mu_{10}\}$) of TAGI-LSTM for both the CB2-3 sensors.

4.2 Forecasting

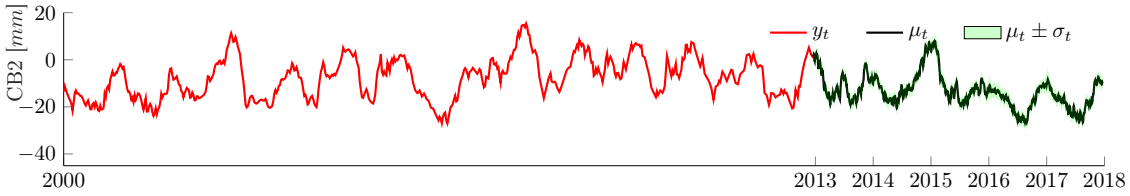
Figure 3 presents the forecasts for the period 2013–2018 obtained from the Gaussian mixture of the BDLM and ten TAGI-LSTM models.

4.3 Model interpretation

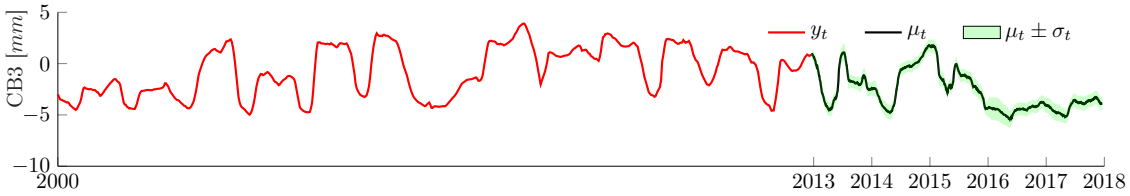
The model interpretation is made using the BDLM model only, as LSTM networks offer little help in understanding the nature of the dependencies between time series.

4.3.1 Dependencies and interaction between time-series

Figures 4 & 5 summarize the information that can be extracted from the BDLM model; (a) presents the relative importance of each component measured by the relative variance of each sub-component; (b) plot the non-linear relationships between the dam's response and the the mean-



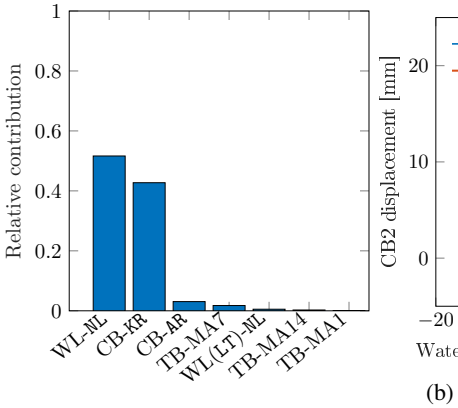
(a) BDLM ∩ 10 TAGI-LSTM instances | CB2



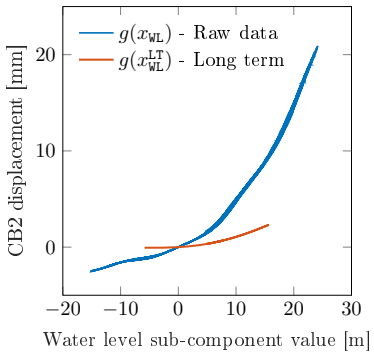
(b) BDLM ∩ 10 TAGI-LSTM instances | CB3

Figure 3. Forecast for the Gaussian mixture made from BDLM forecasts and 10 instances of TAGI-LSTM for both the CB2-3 sensors.

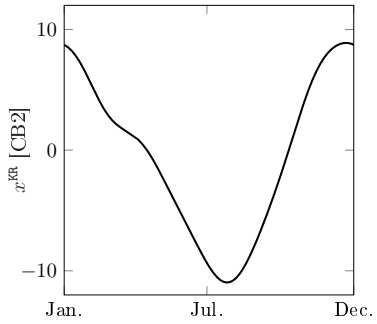
centered water level as well as its long-term pattern as depicted in (d) with corresponding colors; (c) presents the periodic pattern extracted from the CB sensors that can include part of the temperature and water level effects; (d) presents the mean-centered water level as well as the long-term pattern extracted from it by BDLM; (e) presents the model residuals ($x^{\mathtt{AR}}$), i.e., the remaining part of the observation not attributed to observation errors not explained by the other components.



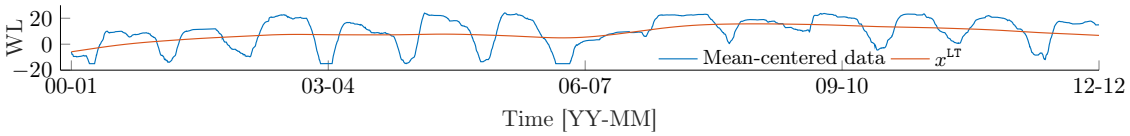
(a) Relative contribution



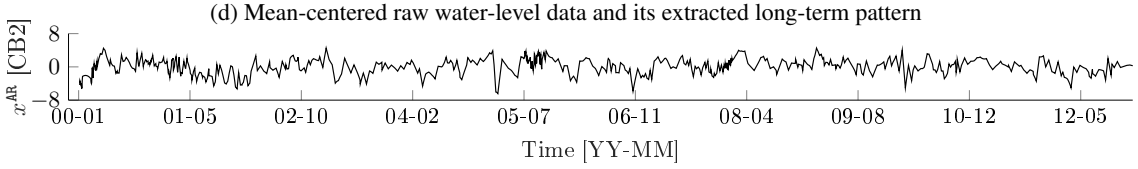
(b) Non-linear dependencies



(c) Periodic component



(d) Mean-centered raw water-level data and its extracted long-term pattern

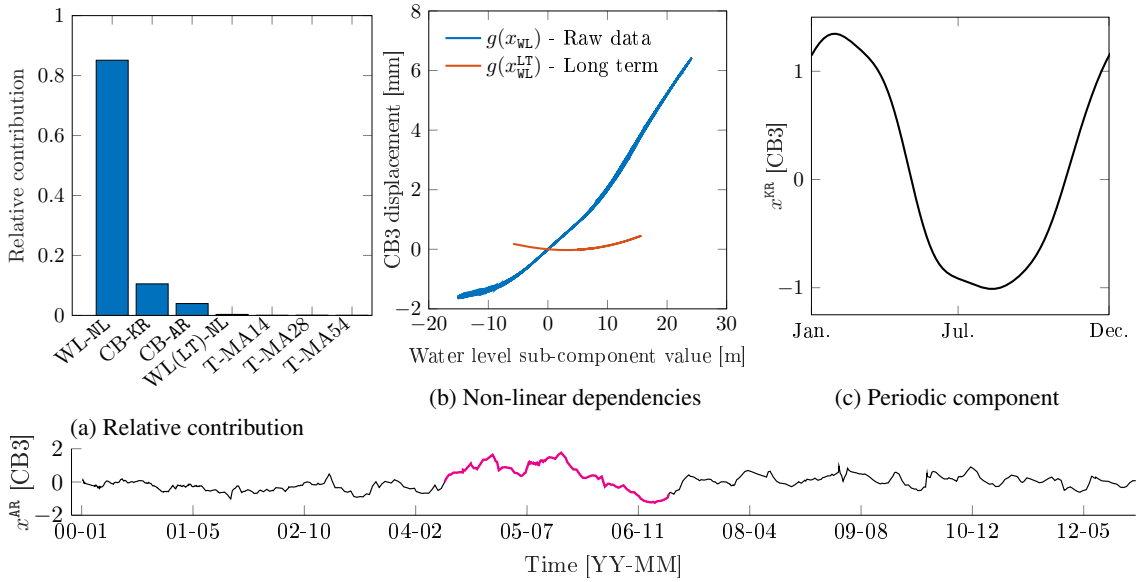


(e) Residual term that cannot be explained for the CB2 sensor

Figure 4. Graphs illustrating the interpretation of the CB2 dataset that can be made from the BDLM components.

For the sensor CB2, we note in Figure 4a, the dominant relative importance of the mean-centered water level through the non-linear dependency $g(x_{\mathtt{WL}})$ depicted in Figure 4b (WL-NL),

and secondly of the periodic pattern $x^{\texttt{KR}}$ depicted in Figure 4c (CB-KR). The third most important contributor is the autoregressive component $x^{\texttt{AR}}$ depicted in Figure 4e (CB-AR), which represents what cannot be explained by the model. Although the relative importance of other components are less than the residual term, they still matter because we are interested in detecting anomalies having small magnitudes. For example, an anomaly with a magnitude of 0.5mm would still have a relative importance comparable to the one day moving average presented in Figure 4a (TB-MA1). Note for instance that the relative importance of the long-term pattern (see 4d) within the water level through the non-linear dependency $g(x^{\texttt{LT}}_{\texttt{WL}})$ depicted in Figure 4b (WL(LT)-NL) is key in order to enable the detection of small anomalies as further detailed in Section 4.3.2.



(a) Relative contribution (b) Non-linear dependencies (c) Periodic component

(d) Residual term that cannot be explained for the CB3 sensor where the non-stationary part is outlined in magenta

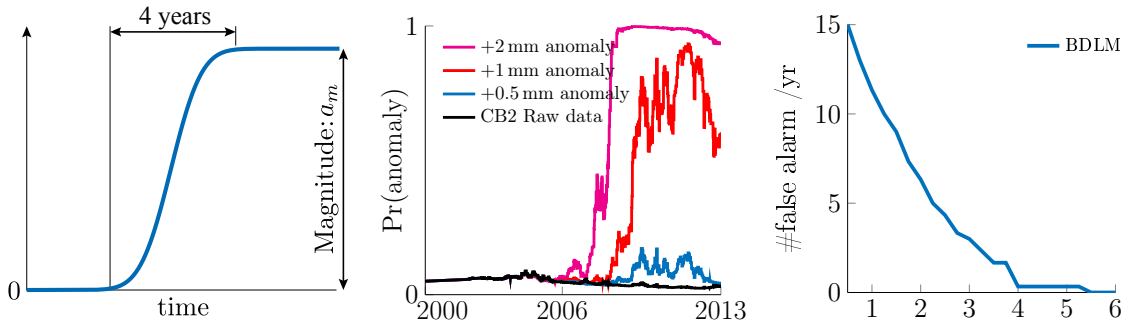
Figure 5. Graph illustrating the interpretation of the CB3 dataset that can be made from the BDLM components.

For the sensor CB3, the contribution of the water level through the non-linear dependency $g(x_{\texttt{WL}})$ depicted in Figure 5b is even more dominant than in the case of CB2. One particularity for CB3 is that the residual term corresponding to the autoregressive component in Figure 5d presents a non-stationary pattern between February 2004 and 2007 as outlined in magenta. The presence of such a pattern indicates that the current components considered in our model for CB3 are missing a part of the dam's behavior. The next section will further investigate this non-stationarity by showing how using a regime-switching analysis would have been able to detect such anomalous occurence in real time.

4.3.2 Anomaly detection using regime switching

As mentioned in §3.1, one key strength of BDLM, is its capacity to detect regime switches (Nguyen & Goulet, 2018a; Khazaeli et al., 2021). We performed such an analysis on the CB2/3 datasets while a first normal regime is modelled with a constant baseline through time, and a second abnormal regime is modelled with a constant-speed regime through time.

For the CB2 sensor, the black curve in Figure 6b presents the probability of anomaly identified using the switching Kalman filter (SKF). This probability close to zero across the dataset indicates that the dam's behaviour is stationary. We further confirm this conclusion by adding to the original data synthetic anomalies of magnitude $a_m = \{0.5, 1, 2\}$ mm building up over a duration of $a_d =$ 4 years, as depicted in Figure 6a. The probability of anomaly identified by the SKF rises in



(a) Schematic representation of synthetic anomalies added on the raw data to quantify anomaly detectability

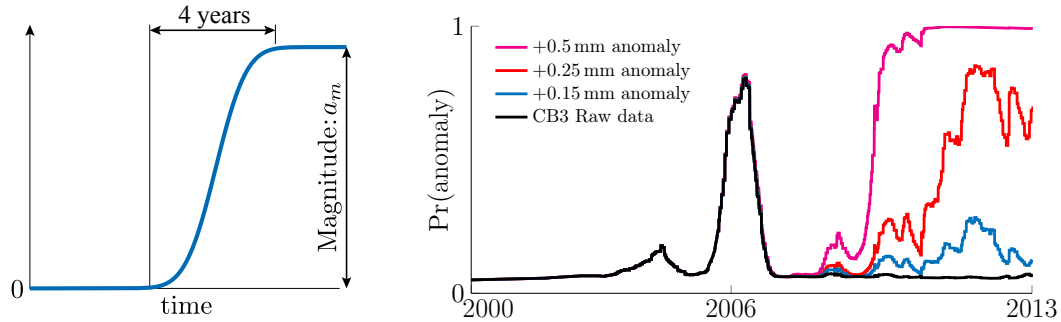
(b) Probability of switching from a stationary to a non-stationary regime taken as a proxy for Pr(anomaly)

(c) Relation on alarm threshold and the number of false alarms

Figure 6. Comparison of the regime switching approach with a threshold-based one for the purpose of detecting anomalies while avoiding false alarms.

all three cases where synthetic anomalies are added, thus correctly indicating their presence. In comparison, if we use an alarm-triggering threshold of 1 mm on the absolute difference between the predicted and measured values for the validation set presented in Figure 2a, we would on average, trigger more than 10 false alarms per year while no alarm should have been triggered. Figure 6c presents the result of this exercice repeated for alarm-triggering thresholds ranging from 0.5 up to 6 mm. Note that these anomaly magnitudes are all smaller than the amplitude of the residual term presented in Figure 4e. This shows that detecting anomalies based on the exceedence of threshold values is prone to false alarms and offers a poor anomaly detectability in comparison with the regime switching approach of Bayesian dynamic linear models. If one chooses a more robust criterion involving multiple successive crossings, the false alarm rate will indeed drop; Nevertheless, this strategy remains poorly suited for the detection of anomalies developing over the span of several years, as parametric models (e.g. HST, LSTM, SVM, etc.) will need to be retrained periodically in order to avoid drift, thus incorporating the changes due to the presence of an anomaly in the updated models,.

Figure 7b presents the same exercise applied to the CB3 sensor. In this case, the SKF detects a regime switch shortly before 2006 as indicated by the jump in the black curve. This regime change can be confirmed by looking at the residual term presented in Figure 5d, where a non-stationary pattern is visually observable between 2004 and 2007. As this pattern disappears after



(a) Schematic representation of synthetic anomalies added on the raw data to quantify anomaly detectability

(b) Probability of switching from a stationary to a non-stationary regime taken as a proxy for Pr(anomaly)

Figure 7. Regime switching analysis applied to the CB3 sensor for the raw data as well as additional synthetic anomalies.

2007 while the probability of regime switch return to 0 after 2006, we speculate that events other than those considered in our model have taken place during that period. We tested our capacity to detect anomalies on CB3 by adding synthetic anomalies as depicted in Figure 7a, with magnitudes $a_m = \{0.15, 0.25, 0.5\}$ mm which are building up over a duration of $a_d = 4$ years. Note that the anomaly has been shifted after 2006 in order not to interfere with the actual anomaly present in the data. We can see in 7b that synthetic anomalies with a magnitude low as 0.15mm are detectable for this sensor.

5 DISCUSSION

The presence of a non-linear residual term for the sensor CB3 lead us to think that, in the context of this benchmark, the long-term predictive capacity for that sensor will be limited. In order to improve the model, it would be worth further investigating (1) the relationship between the anomaly detected on the sensor CB3 and the seepage and piezometric levels measured, (2) the possibility that the long-term effects of the water level on the sensors CB2/3 (see figure 4d) may be related to creep/creep-relief effects (Bažant & Wu, 1974), and (3) following the results of this forecasting competition, if other submissions have identified features explaining the dam's behavior that were not considered here, these could be included in our BDLM model in order to further improve its forecasting accuracy and anomaly detectability.

Despite these limitations, as mentioned in Sections §3.1 & 4.3.2, the key aspect of our method is to enable the detection of anomalies based upon the probability of regime switch rather than on threshold crossing. This has enabled in §4.3.2, the detection of anomalies that are smaller than the residual terms for the CB2 and CB3 sensors. This shows that the anomaly detectability of our method can be decoupled from its long-term prediction capacity.

6 CONCLUSION

This paper presents the results of our submission to the ICOLD's dam prediction benchmark. We provide through our probabilistic ensembling method the explainability of BDLMs as well as the accuracy and ease of use of Bayesian LSTM. Although the benchmark focusses on prediction accuracy and threshold value definition for the purpose of anomaly detection, one should keep in mind that this way of envisioning anomaly detection is only one approach among many others. We showed in this paper that with our probabilistic regime switching method we expect to be able to detect anomalies of 0.5 mm for CB2 and 0.15 mm for CB3, where both can develop over the span of years.

ACKNOWLEDGEMENTS

The first and second authors were financially supported by research grants from Hydro-Quebec/IREQ, and the Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank Vincent Roy and Simon-Nicolas Roth (Hydro-Quebec) for providing insightful comments on the results interpretation and for having revised the manuscript.

REFERENCES

Bažant, Z. P. & Wu, S. (1974). "Rate-type creep law of aging concrete based on maxwell chain." *Matériaux et Construction*, 7(1), 45–60.

Gaudot, I., Nguyen, L. H., Khazaeli, S., & Goulet, J.-A. (2019). "OpenBDLM, an open-source software

for structural health monitoring using bayesian dynamic linear models." *13th Proceedings from the 13th International Conference on Applications of Statistics and Probability in Civil Engineering (ICASP)* (May).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goulet, J.-A. (2017). "Bayesian dynamic linear models for structural health monitoring." *Structural Control and Health Monitoring*, 24(12), 1545–2263.

Goulet, J.-A. (2020). "Chapter 12 - State-Space Models." *Probabilistic Machine Learning for Civil Engineers*, MIT press.

Goulet, J.-A. & Koo, K. (2018). "Empirical validation of bayesian dynamic linear models in the context of structural health monitoring." *Journal of Bridge Engineering*, 23(2), 05017017.

Goulet, J.-A., Nguyen, L., & Amiri, S. (2021). "Tractable approximate gaussian inference for bayesian neural networks." *Journal of Machine Learning Research*, 22(251), 1–23.

Kalman, R. E. (1960). "A new approach to linear filtering and prediction problems." *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D), 35–45.

Khazaeli, S., Nguyen, L., & Goulet, J.-A. (2021). "Anomaly detection using state-space models and reinforcement learning." *Structural Control and Health Monitoring*, 28(6), e2720.

Nguyen, L. H., Gaudot, I., & Goulet, J.-A. (2019). "Uncertainty quantification for model parameters and hidden state variables in bayesian dynamic linear models." *Structural Control and Health Monitoring*, 26(3), e2309 e2309 stc.2309.

Nguyen, L. H. & Goulet, J.-A. (2017). "Structural health monitoring with dependence on hidden non-harmonic covariates." *Submitted to Engineering Structures*.

Nguyen, L. H. & Goulet, J.-A. (2018a). "Anomaly detection with the switching kalman filter for structural health monitoring." *Structural Control and Health Monitoring*, 25(4), e2136.

Nguyen, L. H. & Goulet, J.-A. (2018b). "Structural health monitoring with dependence on non-harmonic periodic hidden covariates." *Engineering Structures*, 166, 187 – 194.

Runnalls, A. R. (2007). "Kullback-leibler approach to gaussian mixture reduction." *IEEE Transactions on Aerospace and Electronic Systems*, 43(3), 989–999.

Sagi, O. & Rokach, L. (2018). "Ensemble learning: A survey." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249.

Salazar, F., Morán, R., Toledo, M., & Oñate, E. (2017). "Data-based models for the prediction of dam behaviour: a review and some methodological considerations." *Archives of computational methods in engineering*, 24(1), 1–21.

APPENDIX

A BDLM MODEL STRUCTURE

The BDLM components used for modeling the independent patterns for CB2/3 are local level (LL), kernel regression (KR) and autoregressive (AR). The mean-centered raw water-level and its long-term pattern (Figure 4d) are modeled using an AR and a local trend (LT) component respectively. The two nonlinear patterns for CB2/3 that are nonlinearly dependent on these two time series are each modeled using a state-regression (SR) component. The moving averages of the temperature (TB) are modeled using AR components. The transition matrices for LL, LT, KR, and AR components (Goulet, 2020, 2017) are as follow:

$$\mathbf{A}_t^{\mathtt{LL}} = 1,\ \mathbf{A}_t^{\mathtt{LT}} = \begin{bmatrix} 1 & \Delta t \\ 0 & 1 \end{bmatrix},\ \mathbf{A}_t^{\mathtt{KR}} = \begin{bmatrix} 0 & \tilde{k}^{\mathtt{KR}}(t, \mathbf{t}^{\mathtt{KR}}) \\ \mathbf{0}_{n\times 1} & \mathbf{I}_n \end{bmatrix},\ \mathbf{A}_t^{\mathtt{AR}} = \phi^{\mathtt{AR}}, \tag{1}$$

where n represents the number of control points for kernel regression and $\Delta t = 1$day. The observation matrices for these components are given by

$$\mathbf{C}_t^{\mathtt{LL}} = 1,\ \mathbf{C}_t^{\mathtt{LT}} = [1\ 0],\ \mathbf{C}_t^{\mathtt{KR}} = [1\ \mathbf{0}_{n\times 1}],\ \mathbf{C}_t^{\mathtt{AR}} = 1. \tag{2}$$

The process noise $\mathbf{Q}_t$ covariance matrices are

$$\mathbf{Q}_t^{\mathtt{LL}} = (\sigma_w^{\mathtt{LL}})^2,\ \mathbf{Q}_t^{\mathtt{LT}} = (\sigma_w^{\mathtt{LT}})^2 \begin{bmatrix} \frac{\Delta t^4}{4} & \frac{\Delta t^3}{2} \\ \frac{\Delta t^3}{2} & \Delta t^2 \end{bmatrix},\ \mathbf{Q}_t^{\mathtt{KR}} = \begin{bmatrix} (\sigma_0^{\mathtt{KR}})^2 & \mathbf{0} \\ \mathbf{0} & (\sigma_1^{\mathtt{KR}})^2 \cdot \mathbf{I}_n \end{bmatrix},\ \mathbf{Q}_t^{\mathtt{AR}} = (\sigma_w^{\mathtt{AR}})^2, \tag{3}$$

The SR component includes $n = 20$ hidden states for the kernel values, $\mathbf{x}^{\mathtt{SK}} = [x_1^{\mathtt{SK}}\ x_2^{\mathtt{SK}}\ \ldots x_n^{\mathtt{SK}}]^\intercal$; $n+1$ hidden states for the regression coefficient that includes n hidden states of control-points, $\mathbf{x}^{\phi^{\mathtt{R}}} = [x_1^{\phi^{\mathtt{R}}}\ x_2^{\phi^{\mathtt{R}}}\ \ldots x_n^{\phi^{\mathtt{R}}}]^\intercal$ and $x_0^{\phi^{\mathtt{R}}}$ which is the hidden state of the predicted regression coefficient; hidden state for the nonlinear pattern for displacement, $x^{\mathtt{S,D}} = \left(x_0^{\phi^{\mathtt{R}}} \cdot x^{\mathtt{S,T}}\right)$ where $x^{\mathtt{S,T}}$ represents the covariate for modeling the nonlinear dependency, and n product terms, $\mathbf{x}^{\mathtt{SKR}} = [x^{\mathtt{SKR},1}\ x^{\mathtt{SKR},2}\ \ldots x^{\mathtt{SKR},n}]^\intercal$, where, $x^{\mathtt{SKR},i} = \left(x_i^{\mathtt{SK}} \cdot x_i^{\phi^{\mathtt{R}}}\right); \forall i = 1:n$. The hidden states for the SR component can be grouped together as

$$\mathbf{x}^{\mathtt{SR}} = [(\mathbf{x}^{\mathtt{SK}})^\intercal\ (\mathbf{x}^{\phi^{\mathtt{R}}})^\intercal\ x_0^{\phi^{\mathtt{R}}}\ x^{\mathtt{S,D}}\ (\mathbf{x}^{\mathtt{SKR}})^\intercal]^\intercal.$$

The transition matrix for the SR component of size $3n+2$ is formulated as

$$\mathbf{A}_t^{\mathtt{SR}} = \begin{bmatrix} \mathbf{0}_n & 0_{1\times n} & 0 & 0 & 0_{1\times n} \\ \vdots & \mathbf{I}_n & 0 & 0 & 0_{1\times n} \\ \vdots & \ldots & 0 & 0 & \mathbf{1}_{1\times n} \\ \vdots & \ldots & \ldots & 0 & \mathbf{0}_{1\times n} \\ sym. & \ldots & \ldots & \ldots & \mathbf{0}_n \end{bmatrix}. \tag{4}$$

The observation matrix $\mathbf{C}_t^{\mathtt{SR}}$ is given by

$$\mathbf{C}_t^{\mathtt{SR}} = \left[\mathbf{0}_{n\times 1}^\intercal\ \mathbf{0}_{n\times 1}^\intercal\ 0\ 1\ \mathbf{0}_{n\times 1}^\intercal\right]. \tag{5}$$

No process noise is considered for the SR component and is given by $\mathbf{Q}_t^{\mathtt{SR}} = \mathbf{0}_{3n+2}$. Using equations 1 & 4, the global transition matrix $\mathbf{A}_t$ is obtained by arranging the transition matrices block diagonally shown by

$$\mathbf{A}_t = \text{blockdiag}\left(\overbrace{[\mathbf{A}_t^{\mathtt{LL}},\ \mathbf{A}_t^{\mathtt{KR}},\ \mathbf{A}_t^{\mathtt{AR}}]}^{\text{CB2/3}}, \overbrace{[\mathbf{A}_t^{\mathtt{LT}},\ \mathbf{A}_t^{\mathtt{SR}_1}]}^{\text{WL1}}, \overbrace{[\mathbf{A}_t^{\mathtt{AR}},\ \mathbf{A}_t^{\mathtt{SR}_2}]}^{\text{WL2}}, \overbrace{[\mathbf{A}_t^{\mathtt{AR}}]}^{\text{T-MA1}}, \overbrace{[\mathbf{A}_t^{\mathtt{AR}}]}^{\text{T-MA7}}, \overbrace{[\mathbf{A}_t^{\mathtt{AR}}]}^{\text{T-MA14}}, \overbrace{[\mathbf{A}_t^{\mathtt{AR}}]}^{\text{T-MA28}}, \overbrace{[\mathbf{A}_t^{\mathtt{AR}}]}^{\text{T-MA54}}\right), \tag{6}$$

where WL1 and WL2 refers to the long term pattern and the mean-centered raw water level, and the nonlinear dependencies are modeled using the $\mathtt{SR}_1$ and $\mathtt{SR}_2$ components. Using equations 2 & 5, the global observation matrix $\mathbf{C}_t$ is given by

$$\mathbf{C}_t = \text{blockdiag}\left(\overbrace{[\mathbf{C}_t^{\mathtt{LL}}, \mathbf{C}_t^{\mathtt{KR}}, \mathbf{C}_t^{\mathtt{AR}}]}^{\text{CB2/3}}, \overbrace{[\mathbf{C}_t^{\mathtt{LT}}, \mathbf{C}_t^{\mathtt{SR}_1}]}^{\text{WL1}}, \overbrace{[\mathbf{C}_t^{\mathtt{AR}}, \mathbf{C}_t^{\mathtt{SR}_2}]}^{\text{WL2}}, \overbrace{[\mathbf{C}_t^{\mathtt{AR}}]}^{\text{T-MA1}}, \overbrace{[\mathbf{C}_t^{\mathtt{AR}}]}^{\text{T-MA7}}, \overbrace{[\mathbf{C}_t^{\mathtt{AR}}]}^{\text{T-MA14}}, \overbrace{[\mathbf{C}_t^{\mathtt{AR}}]}^{\text{T-MA28}}, \overbrace{[\mathbf{C}_t^{\mathtt{AR}}]}^{\text{T-MA54}}\right). \quad (7)$$

The $\mathbf{Q}_t$ and the $\mathbf{R}_t$ matrices are

$$\mathbf{Q}_t = \text{blockdiag}\left(\overbrace{[\mathbf{Q}_t^{\mathtt{LL}}, \mathbf{Q}_t^{\mathtt{KR}}, \mathbf{Q}_t^{\mathtt{AR}}]}^{\text{CB2/3}}, \overbrace{[\mathbf{Q}_t^{\mathtt{LT}}, \mathbf{Q}_t^{\mathtt{SR}_1}]}^{\text{WL1}}, \overbrace{[\mathbf{Q}_t^{\mathtt{AR}}, \mathbf{Q}_t^{\mathtt{SR}_2}]}^{\text{WL2}}, \overbrace{[\mathbf{Q}_t^{\mathtt{AR}}]}^{\text{T-MA1}}, \overbrace{[\mathbf{Q}_t^{\mathtt{AR}}]}^{\text{T-MA7}}, \overbrace{[\mathbf{Q}_t^{\mathtt{AR}}]}^{\text{T-MA14}}, \overbrace{[\mathbf{Q}_t^{\mathtt{AR}}]}^{\text{T-MA28}}, \overbrace{[\mathbf{Q}_t^{\mathtt{AR}}]}^{\text{T-MA54}}\right), \quad (8)$$

$$\mathbf{R}_t = \text{blockdiag}\left(\overbrace{[(\sigma_{v_1})^2}^{\text{CB2/3}}, \overbrace{(\sigma_{v_2})^2}^{\text{WL1}}, \overbrace{(\sigma_{v_3})^2}^{\text{WL2}}, \overbrace{(\sigma_{v_4})^2}^{\text{T-MA1}}, \overbrace{(\sigma_{v_5})^2}^{\text{T-MA7}}, \overbrace{(\sigma_{v_6})^2}^{\text{T-MA14}}, \overbrace{(\sigma_{v_7})^2}^{\text{T-MA28}}, \overbrace{(\sigma_{v_8})^2]}^{\text{T-MA54}}\right), \quad (9)$$

where σ_{v_i}, $\forall i = 1:8$ refers to the standard deviation of the observation noise for each of the time series.

B LSTM MODEL STRUCTURE

We use two separate models which have the same architecture, but do not share the parameters to predict the displacements CB2 and CB3. The common network's architecture consists of one LSTM layer of 50 hidden units, and a dense layer with no activation function to combine the output of the LSTM layer. In order to take into account the lagging effect that the temperature and the reservoir's level may have on the displacement, we use a sequence of length $\mathtt{M}$ of covariates as inputs for the LSTM instead of using only the covariates at time t as

$$\boldsymbol{h}_t = g(\boldsymbol{h}_{t-1}, y_{t-\mathtt{L}:t-1}, \boldsymbol{x}_{t-\mathtt{M}+1:t}),$$

where $g(\cdot)$ is the function implemented by the LSTM, $\boldsymbol{h}$ are the hidden states, y is the displacement observation, $\boldsymbol{x}$ are covariates which are the reservoir's level and temperature TB, and $\mathtt{L}$ is the lookback period. During training when the data is missing, and during prediction when the data is not available, we replace y by the network's prediction, and $\boldsymbol{x}$ by $\mathbf{0}$. When using TAGI to perform Bayesian estimation for the LSTM network parameters, it is required to define an observation noise for each time series (Goulet et al., 2021). The standard deviation for this observation noise is a hyper-parameter which needs to be learnt from data. We perform a grid-search to find the best hyper-parameter values for each model as presented in Table 1. For each candidate value in the grids, we train our models with early-stopping on a subset of training data from 2000 to end of 2009, and report the log-likelihood for the validation period from 2010 to end of 2012. The values which maximize the log-likelihood of the validation set are chosen as the final hyper-parameter values.

Table 1. LSTM hyper-parameters

Hyper-parameter	CB2	CB3	Grid
Observation noise's standard deviation	0.05	0.01	$\{0.01, 0.05, 0.1, 0.15\}$
$\mathtt{L}$	35	14	$\{14, 35, 56, 70\}$
$\mathtt{M}$	21	21	$\{7, 21, 35, 49, 70\}$

C LOG-LIKELIHOOD AND WEIGHT

Table 2. Log-likelihood measured between 2012-2013 and weight by each model

Model	CB2		CB3	
	Log-likelihood	w_i	Log-likelihood	w_i
BDLM	−63.89	0.506	−12.06	0.659
LSTM #1	−66.48	0.049	−46.96	0.017
LSTM #2	−66.53	0.049	−48.66	0.016
LSTM #3	−64.74	0.050	−11.64	0.068
LSTM #4	−64.98	0.050	−45.52	0.017
LSTM #5	−73.79	0.044	−35.36	0.022
LSTM #6	−65.14	0.050	−37.59	0.021
LSTM #7	−66.12	0.049	−24.71	0.032
LSTM #8	−62.4	0.052	−33.49	0.023
LSTM #9	−63.18	0.051	−28.68	0.028
LSTM #10	−62.51	0.052	−8.37	0.095

D MEAN ABSOLUTE ERROR (MAE)

Table 3. MAE for the validation period between 2010-2013

Model	CB2	CB3
Mixture	1.366	0.253
BDLM	1.312	0.248
LSTM #1	1.574	0.490
LSTM #2	2.109	0.537
LSTM #3	1.910	0.486
LSTM #4	2.008	0.534
LSTM #5	1.945	0.610
LSTM #6	1.833	0.566
LSTM #7	1.836	0.393
LSTM #8	1.705	0.524
LSTM #9	1.796	0.452
LSTM #10	1.869	0.462