# Gaussian Process Regression for Monitoring a Secondary Settler

**Jesús Zambrano[1], Oscar Samuelsson[1,2], Tatiana Chistiakova[1], Hongbin Liu[1], Bengt Carlsson[1]**

[1]Division of System and Control, Department of Information Technology, Uppsala University, P.O. Box 337, 75105 Uppsala, Sweden. (E-mail: jesus.zambrano@it.uu.se; tatiana.chistiakova@it.uu.se; hongbin.liu@it.uu,se; bengt.carlsson@it.uu.se).
[2]IVL Swedish Environmental Research Institute, P.O. Box 210 60, 10031 Stockholm, Sweden. (E-mail: oscar.samuelsson@ivl.se).

**Abstract:** An approach based on Gaussian Process Regression for monitoring the sludge profile of a secondary settler is proposed. Gaussian Process is a probabilistic, nonparametric model with an uncertainty prediction. The approach is illustrated using data from a sensor measuring the sludge concentration in a settler as a function of the settler level at Bromma wastewater treatment plant (WWTP). Results suggest that the approach is feasible for monitoring and fault detection of the sludge settling process.

**Keywords:** covariance function; fault detection; Gaussian process regression; monitoring; secondary settler; sludge profile.

## INTRODUCTION

Increasing demands on effluent water quality and resource efficient operation are important driving forces for wastewater treatment plants (WWTP's). Process monitoring and detection of abnormal process conditions are important tools to secure a robust and efficient process. Furthermore, an increased amount of sensors, adding process information but also complexity for plant operators, contributes to the need of fault detection methods.

Sedimentation is one of the most important processes which determines the performance of the activated sludge process (ASP), nevertheless the performance of secondary settling tanks (SST's) is often far from satisfactory (Li and Stenstrom 2014). A great effort has been put in to understand, model and predict the settling behavior. The current knowledge about one-dimensional settling models can be found in a recent review by (Li and Stenstrom 2014) and the references therein. Despite increased knowledge in settling behavior, the number of full-scale applications to monitor the condition of the SST is limited. Some examples of existing monitoring methods for SST performance include methods based on: image analysis (Grijspeerdt and Verstraete 1997) and model-based approaches (Traoré et al. 2006, Yoo et al. 2002).

One alternative to previously mentioned monitoring methods would be a data-based approach. Here, Gaussian process regression (GPR) is one possible technique. GPR is a non-parametric regression method where a prediction of the response variable is given as a probability density function. Thus, the predicted value of the response variable comes with a variance estimate, which is interpreted as an uncertainty measure of the prediction. The method is thoroughly described in (Rasmussen and Williams 2005) and has gained large interest within the machine learning community, and more recently within systems identification (Chen et al. 2012). It is worth to note that GPR is not a new concept, it was originally known as Kriging, with an origin from geostatistics in the 1950s (Cressie 1990).

GPR has several properties making it useful for fault detection, such as: probabilistic prediction including an uncertainty estimate, flexible regression in a non-parametric fashion, and it is relatively simple to implement in common programming languages. Although GPR is a non-parametric method, an appropriate covariance function, has to be selected by the user. The covariance function determines the model structure of the GPR and prior knowledge can be

included. It is an active area of research of how to construct good covariance functions, see e.g. (Lloyd et al. 2014) for an automated selection approach.

GPR has been used for fault detection in a wide area of applications (Roberts et al. 2013). Some examples are: maritime vessel track analysis (Smith et al. 2012), change point detection (Garnett et al. 2010), bearings (Boškoski et al. 2015) and process monitoring (Serradilla et al. 2011). GPR has also been used in various environmental applications such as: monitoring of water quality sensors (Osborne et al. 2012), modelling of an anaerobic wastewater treatment system (Ni et al. 2012), modelling nitrification process and biomass growth (Ažman and Kocijan 2007), uncertainty analyses of an anaerobic digestion model (Južnič-Zonta et al. 2012) and for control of an SBR-reactor (Kocijan and Hvala 2013).

The objective of this study is to present a GPR-based approach for monitoring the sludge profile of a secondary settler in an activated sludge process. The aim is to automatically detect deviations from normal conditions. The main concept of this approach is to use the GPR methodology to obtain a non-faulty zone, where the mapping of new profiles is evaluated. Hence, this mapping is used to decide if the new profile belongs to non-faulty or faulty condition.

The paper is organized as follows. First, an introduction to Gaussian Process Regression is detailed, which includes the fault detection criteria based on the GP implementation. Later, a case study showing a practical application of the GP-approach is presented. Next, results and discussions are included. Finally, conclusions are drawn.

## GAUSSIAN PROCESS REGRESSION

A Gaussian Process (GP) is a collection of random variables which has a joint Gaussian distribution. Assume we observe some inputs $x_i$ and some outputs $y_i$ from a certain process, and that $y_i = f(x_i)$. The optimal approach is to infer a distribution over functions given the data. A GP is completely specified by its mean $m(x_i)$ and covariance function $k(x_i, x_j)$, and a distribution over the functions $f(x_i)$ can therefore be expressed as

$$f(x_i) \sim GP\left(m(x_i), k(x_i, x_j)\right) \tag{1}$$

The mean and covariance functions involve a vector of parameters (called hyperparameters) required for the model. The simplest approach to optimize the hyperparameters is to maximize the log-likelihood function of the dataset, see more details in (Rasmussen and Williams 2005).

$$\ln[p(\boldsymbol{Y}|\boldsymbol{X})] = -\frac{1}{2}\boldsymbol{Y}^T(K + {\sigma_n}^2 I)^{-1}\boldsymbol{Y} - \frac{1}{2}\ln(|K + {\sigma_n}^2 I|) - \frac{N}{2}\ln(2\pi) \tag{2}$$

where $\boldsymbol{X} = [x_1, \dots, x_N]$ and $\boldsymbol{Y} = [y_1, \dots, y_N]^T$ are $N$ observed data with Gaussian noise of variance ${\sigma_n}^2$, $K = k(\boldsymbol{X}, \boldsymbol{X})$ is a $N \times N$ covariance matrix of the training dataset, $I$ is a $N \times N$ identity matrix.

A regression in a GP means that, based on the given data set $D = (\boldsymbol{X}, \boldsymbol{Y})$, and a new input $x_*$, we wish to find the predictive distribution of the associated output $y_*$. The predictive distribution of $y_*$ over $D$ is Gaussian described by

$$p(y_*|(\boldsymbol{X}, \boldsymbol{Y}), x_*) = \mathcal{N}(m_*(x_*), {\sigma_*}^2(x_*)) \tag{3}$$

with mean $m_*(x_*)$ and covariance ${\sigma_*}^2(x_*)$ (Rasmussen and Williams 2005)

$$m_*(x_*) = {\boldsymbol{k}_*}^T(K + {\sigma_n}^2 I)^{-1}\boldsymbol{Y} \tag{4}$$

$${\sigma_*}^2(x_*) = k_{**} - {\boldsymbol{k}_*}^T(K + {\sigma_n}^2 I)^{-1}\boldsymbol{k}_* \tag{5}$$

where $\boldsymbol{k}_* = [k(x_1, x_*), \dots, k(x_N, x_*)]^T$ is a $N \times 1$ vector of covariance between the test and the training dataset, $k_{**} = k(x_*, x_*)$ is the auto-covariance of the test dataset.

For all calculations, MATLAB and GPML-toolbox (Rasmussen and Nickisch 2010) have been used.

**Fault detection criteria**

The implementation of GP involves a residual calculation $r_{GP}$. This residual is used to monitor and identify possible faulty conditions in the process. We assume that $r_{GP}$ belongs to one out of two different hypotheses: $H_0$ and $H_1$. The problem can be expressed by the classical binary hypothesis testing problem

$$\begin{aligned} H_0&: r_{GP} \leq h \\ H_1&: r_{GP} > h \end{aligned} \tag{6}$$

where $H_0$ is the non-faulty condition hypothesis and $H_1$ is the faulty condition hypothesis, $h$ is a predefined threshold. The aim is to decide if the system has changed between $H_0$ and $H_1$ when changes in the dynamic of the process are presented. It is assumed that $H_0$ and $H_1$ are equally likely.

For a given group of profiles, we propose the following steps to compute the residual $r_{GP}$:

Step 1: Collect profiles in non-faulty condition (training dataset $(\boldsymbol{X}, \boldsymbol{Y})$).

Step 2: Select a covariance function and determine the hyperparameters by maximizing expression (2).

Step 3: Obtain the predictive distribution $p(y_*|(X, Y), x_*)$ as described by expressions (3) to (5).

Step 4: For a new $j^{th}$ profile formed by $\boldsymbol{X}_* = [x_{*1}, \dots, x_{*i}, \dots, x_{*N}]$ and $\boldsymbol{Y}_* = [y_{*1}, \dots, y_{*i}, \dots, y_{*N}]$ compute:

$$r_{GP}(j) = \frac{1}{N}\sum_{i=1}^{N} v(i) \;\; ; \text{ where } \; v(i) = \begin{cases} 1 & \text{if } |y_{*i} - m_*(x_{*i})| > 2\sigma_*(x_{*i}) \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

A fault is decided if $r_{GP}(j) > h$, where the threshold $h = \max\{r_{GP}\}|_{t \in H_0}$ is the maximum $r_{GP}$ obtained during the training dataset.

Note that the predictive distribution is then used for mapping the new profile. Then, the more the data in the new profile is outside the predictive distribution, the larger the residual $r_{GP}$ will be.

**CASE STUDY: MONITORING A SECONDARY SETTLER**

The approach is tested using real data from a sensor installed in a secondary settler at Bromma WWTP in Stockholm, Sweden. The sensor measures the suspended solids (SS) as a function of the settler level. As shown in Figure 1a, the sensor goes from top to bottom of the settler measuring the level [m] and the SS concentration [g/L]. The profile obtained is called *sludge profile*. A typical sludge profile is shown in Figure 1b.

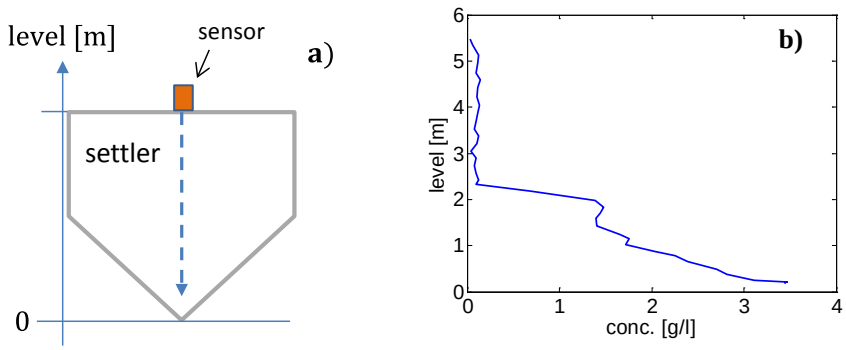


**Figure 1**. **a)** Experiment setup; **b)** Typical sludge profile plotted as level vs concentration.

The sensor works discontinuously, which means that a new sludge profile is automatically measured after a certain period of time (in minutes). The data contained in the sludge profile can be affected by different reasons, including: changes in the return and/or excess of sludge flow rates, sludge scape, large variations in the influent flow and composition and sensor clogging. In this study, the aim is to detect events such as sludge escape and sensor clogging.

As part of the experiment, two additional measurements were recorded: the level at which the SS concentration is equal to 0.5g/L (called *fluff level*) and 2.5g/L (called *sludge level*). We will refer to these levels during the results and discussions of the experiment.

For the Gaussian distribution, we choose $x$ representing the level of the sensor and $y = f(x)$ representing the SS concentration. The covariance function suitable for our case study was

$$k(x_i, x_j) = (ax_i + b\,) + \exp\left[-\beta_0(x_i - x_j)^2\right] \qquad (8)$$

where the unknown hyperparameters $(a, b, \beta_0)$ are determined from the training dataset. The function in (8) is a composition of a linear function and a squared exponential function.

**RESULTS**

Figure 2a shows several profiles in non-faulty conditions used as training dataset. Figure 2b shows the predictive mean value (red line) along with $\pm 2\sigma_*$ (the predictive distribution of the standard deviation) given by the regression in the GP, see expressions (4) and (5).

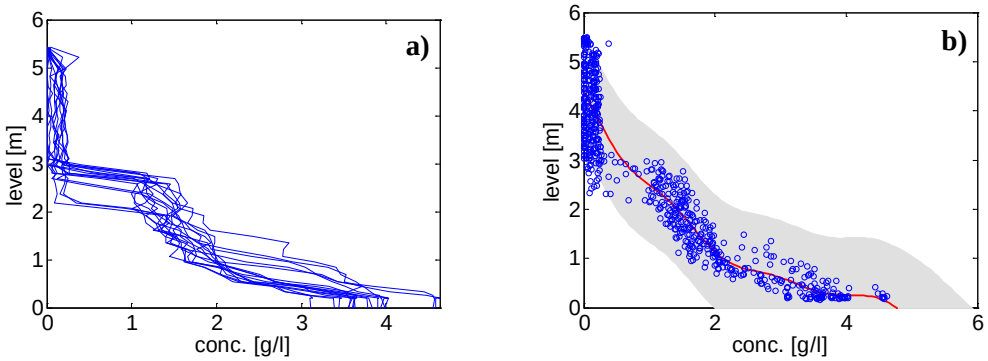


**Figure 2. a)** Sludge profiles (blue lines) used as training dataset $D$; **b)** Predictive distribution over the training dataset $D$ (plotted using blue dots), showing $m_*$ (red line) $\pm 2\sigma_*$ (grey zone).

A total of 17 sludge profiles in non-faulty condition were used as training dataset. Observe from Figure 2b that the predictive distribution provides an interpolation between the training dataset. Once the predictive distribution is obtained, the monitoring of a new sludge profile is feasible, as it was described by the Step 4 for computing the residual profile $r_{GP}$, see expression (7).

Several trials were performed to validate the approach, each of them formed by several sludge profiles. As illustration, we present one trial which consisted of 33 days of settler monitoring. A new sludge profile was measured every 15 minutes, giving a total of 3168 sludge profiles. In order to see the evolution of the profiles during time, these are shown after 10, 20 and 30 days of the experiment, as depicted in Figure 3a to Figure 3c, respectively.

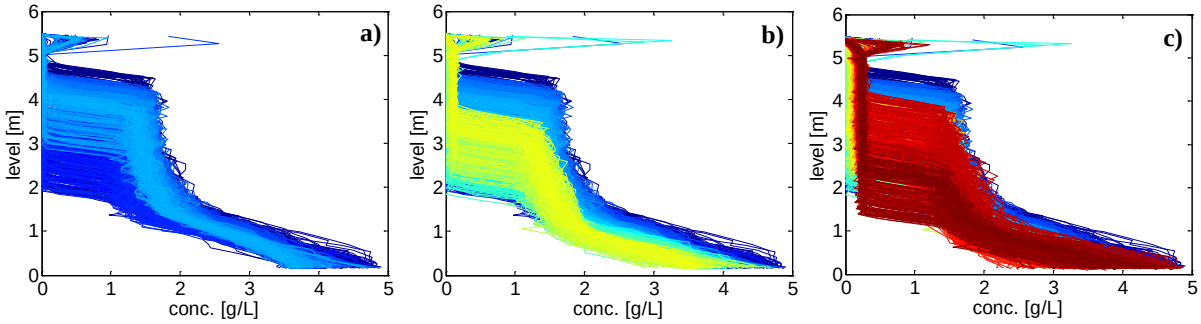


**Figure 3**. Total of profiles scanned: **a)** after 10 days; **b)** after 20 days; **c)** after 30 days.

Figure 4 shows the profile of the GP residual $r_{GP}$ for the experiment, and for comparison, both the fluff level and the sludge level profiles are also plotted.

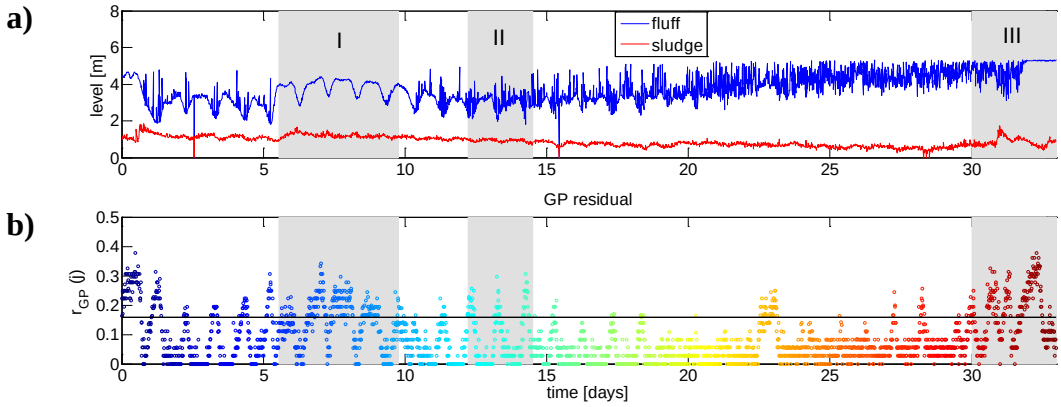


**Figure 4**. **a)** Fluff level (blue line) and sludge level (red line); **b)** GP residual $r_{GP}$ profile and threshold $h$ (horizontal black line).

The residual profile $r_{GP}$ is coloured from dark blue (beginning of experiment) to dark red (end of experiment), which correspond to the same range of colours assigned to the sludge profiles in Figure 3.

Figure 4 shows some abnormal behaviours obtained during the experiment, marked as Period I to III. In particular, Period I refers to large variation in the influent flow rate, causing fluctuations in the sludge blanket, this effect can also be seen in the variation of the fluff level profile, see the sludge profiles of this event in Figure 5a. Note also some peaks in the $r_{GP}$ occurring in Period II, which were due to abrupt changes in the sludge profile at the beginning of the measurement, see Figure 5b. Another type of event was the sensor clogging, which started to be detected for profiles in Period III. This event was confirmed by ocular inspection of the sensor and the existence of floating sludge at the surface level, promoting sludge escape, see the sludge profiles of this event in Figure 5c.

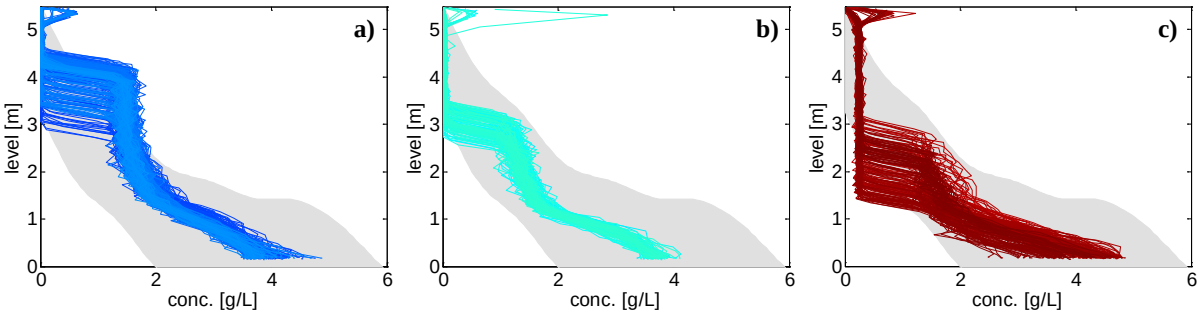


**Figure 5.** Sludge profiles for different periods. **a)** Period I**; b)** Period II**; c)** Period III**.** The predictive distribution of the training dataset (grey zone) is also shown.

## DISCUSSIONS

As shown previously, the predictive distribution is a key factor in the GP-based approach. An important aspect that determines the predictive distribution of the process is the choice of the covariance function. The covariance function used in this study (see expression (8)) is formed by two widely used functions, which based on the predictive distribution obtained (see Figure 2b), this function was suitable for this particular case study. The mean function, first part of expression (8), was required to obtain a negative slope of the sludge profile, i.e. a non-stationary process. The remaining part of (8), the exponential covariance function, is widely used in GPs and can be seen as a smoothing function. For a different process, giving a different profile, it may require a new definition of the covariance function. Examples concerning the choice of covariance functions to different kind of datasets can be found in (Lloyd et al. 2014) and in (Wilson 2013).

The approach shown in this study can be used to improve the data quality of the sensor, e.g. when a profile includes *outliers*. An outlier can be defined as a sharp change in the measured value between two successive data. In our case study, a sludge profile with outliers means that this data is far from the predictive distribution (grey zone in Figure 2b). This results in a GP residual $r_{GP}$ larger than the threshold. For our case study, the management of outliers was not relevant. For a process where this situation is recurrent, the reconstruction of the profile can be performed by doing a mapping from the predictive distribution. Then, by defining, for example, a region larger that $\pm 4\sigma_*$ or than $\pm 5\sigma_*$ as the region for outliers, then this faulty data is replaced by the predictive mean $m_*(x_*)$, as shown in expression (4).

Another example of improving data quality is when the sludge profile has missing data, i.e. when the amount of data in a profile is incomplete. In our case study we did not deal with this situation but the GP approach can be applied to solve this problem. Similar to the case of the outliers, the reconstruction of the missing data can be performed by mapping the new value from the predictive mean. Naturally, if a sludge profile has several missing data, an alarm must be decided. The number of allowed missing data must be previously defined.

The implementation of the GP-based approach will be determined by a proper definition of the predictive distribution of the process. During the GP design, an approximate idea of the shape of this predictive function will help in the definition of the predictive function.

It is important to remark that given two single sensor measuring the same process, both of them will have different training datasets, resulting that each sensor will have a *unique* predictive distribution for monitoring and fault detection. This gives a remarkable advantage of the developed methodology, because it is a general methodology and not only applicable to data from a specific sensor or process.

Apart from monitoring and fault detection, a possible application of this approach is to use the information given by the residual profile $r_{GP}$ as a control action. In this way, the final goal is to perform the reliability of the settler and, as consequence, the performance of the WWTP.

## CONCLUSIONS

A GP-based approach for monitoring and fault detection of sludge profiles of a secondary settler in a wastewater treatment plant is presented. From a set of non-faulty profiles, the main idea is to obtain a non-faulty zone by means of the GPR methodology. With the aid of this zone, the mapping of a new profile can be evaluated and possible abnormal profiles can be detected. As a practical example, real data was used. Results suggest that this approach can be a valuable tool for monitoring the performance of the settler.

### ACKNOWLEDGMENTS

The authors acknowledge funding support under the European Union's Seventh Framework Programme managed by the Research Executive Agency (REA) http://ec.europa.eu/research/rea (FP7/2007_2013), Grant

Agreement N.315145 (Diamond). Funding from Käppala Association, Syvab and Stockholm Water Company, Foundation for IVL Swedish Environmental Research Institute and the Swedish Water and Wastewater Association is gratefully acknowledged.

**REFERENCES**

Ažman, K. and Kocijan, J. (2007) Application of Gaussian processes for black-box modelling of biosystems. ISA Transactions 46(4), 443-457.

Boškoski, P., Gašperin, M., Petelin, D. and Juričić, Đ. (2015) Bearing fault prognostics using Rényi entropy based features and Gaussian process models. Mechanical Systems and Signal Processing 52–53(0), 327-337.

Chen, T., Ohlsson, H. and Ljung, L. (2012) On the estimation of transfer functions, regularizations and Gaussian processes—Revisited. Automatica 48(8), 1525-1535.

Cressie, N. (1990) The origins of kriging. Mathematical Geology 22(3), 239-252.

Garnett, R., Osborne, M.A., Reece, S., Rogers, A. and Roberts, S.J. (2010) Sequential Bayesian Prediction in the Presence of Changepoints and Faults. The Computer Journal 53(9), 1430-1446.

Grijspeerdt, K. and Verstraete, W. (1997) Image analysis to estimate the settleability and concentration of activated sludge. Water Research 31(5), 1126-1134.

Južnič-Zonta, Ž., Kocijan, J., Flotats, X. and Vrečko, D. (2012) Multi-criteria analyses of wastewater treatment bio-processes under an uncertainty and a multiplicity of steady states. Water Research 46(18), 6121-6131.

Kocijan, J. and Hvala, N. (2013) Sequencing batch-reactor control using Gaussian-process models. Bioresource Technology 137(0), 340-348.

Li, B. and Stenstrom, M.K. (2014) Research advances and challenges in one-dimensional modeling of secondary settling Tanks – A critical review. Water Research 65(0), 40-63.

Lloyd, J.R., Duvenaud, D., Grosse, R., Tenenbaum, J.B. and Ghahramani, Z. (2014) Automatic Construction and Natural-Language Description of Nonparametric Regression Models.

Ni, W., Wang, K., Chen, T., Ng, W.J. and Tan, S.K. (2012) GPR model with signal preprocessing and bias update for dynamic processes modeling. Control Engineering Practice 20(12), 1281-1292.

Osborne, M.A., Garnett, R., Swersky, K. and De Freitas, N. (2012), pp. 349-355.

Rasmussen, C.E. and Nickisch, H. (2010) Gaussian Processes for Machine Learning (GPML) Toolbox. J. Mach. Learn. Res. 11, 3011-3015.

Rasmussen, C.E. and Williams, C.K.I. (2005) Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N. and Aigrain, S. (2013) Gaussian processes for time-series modelling.

Serradilla, J., Shi, J.Q. and Morris, A.J. (2011) Fault detection based on Gaussian process latent variable models. Chemometrics and Intelligent Laboratory Systems 109(1), 9-21.

Smith, M., Reece, S., Roberts, S. and Rezek, I. (2012) Online Maritime Abnormality Detection Using Gaussian Processes and Extreme Value Theory, pp. 645-654, IEEE.

Traoré, A., Grieu, S., Thiery, F., Polit, M. and Colprim, J. (2006) Control of sludge height in a secondary settler using fuzzy algorithms. Computers & Chemical Engineering 30(8), 1235-1242.

Wilson, A.G. (2013) Gaussian process kernels for pattern discovery and extrapolation, pp. 2104-2112.

Yoo, C., Choi, S. and Lee, I.-B. (2002) Adaptive modeling and classification of the secondary settling tank. Korean Journal of Chemical Engineering 19(3), 377-382.