Adaptive Data-driven Modelling and Forecasting of Effluent Treatment Plants

Rihab Abdul Razak,a\* Arvind Ravi,a Resmi Suresh,a Koen de Leeuw,b Jose M Gonzalezb

aShell India Markets Pvt Ltd, Bengaluru, India

bShell Global Solutions International B.V., Amsterdam, The Netherlands

Rihab.AbdulRazak@shell.com

Abstract

Contaminants from industrial wastewater can be removed by biological treatment which is a slow process (residence time of the order of days) with non-linear dynamics. Modelling of its key performance indicators (KPIs) is challenging due to multiple factors such as uncertainties in the upstream processes, environmental conditions and reliability of process and lab measurements. In this paper, we propose a reliable dynamic model that can be used for forecasting the biotreater KPIs. Since developing a physics-based model is challenging in an industrial setup, we propose a data-driven adaptive linear model which can capture the correlations and delays between different variables of interest accurately. We use forecasting methods from time-series analysis to forecast the exogenous inputs. Uncertainties associated with the predictions are also computed.

**Keywords**: Data-driven dynamic model, Forecasting, Biotreater, Effluent treatment

* 1. Introduction

Effluent from industrial processes often contains high quantities of organic chemical compounds with potential harmful effects on the environment (Sunita et al., 2020). The generated effluent stream can be biologically treated and the process water recovered for reuse or discharge. The effluent treatment plant (ETP) consists of large basins or biotreaters where the continuous flow of effluent is brought into contact with microorganisms (biomass). The biotreater is aerated from the bottom to provide dissolved oxygen (DO) for the aerobic microbes to consume the complex organic compounds (Eckenfelder and Musterman, 1998). The microbial activity follows a non-linear dynamics with relatively slow reaction rates (Newhart et al., 2019). Hence, long residence time is established in these biotreaters to reach the desired degradation of organic compounds.

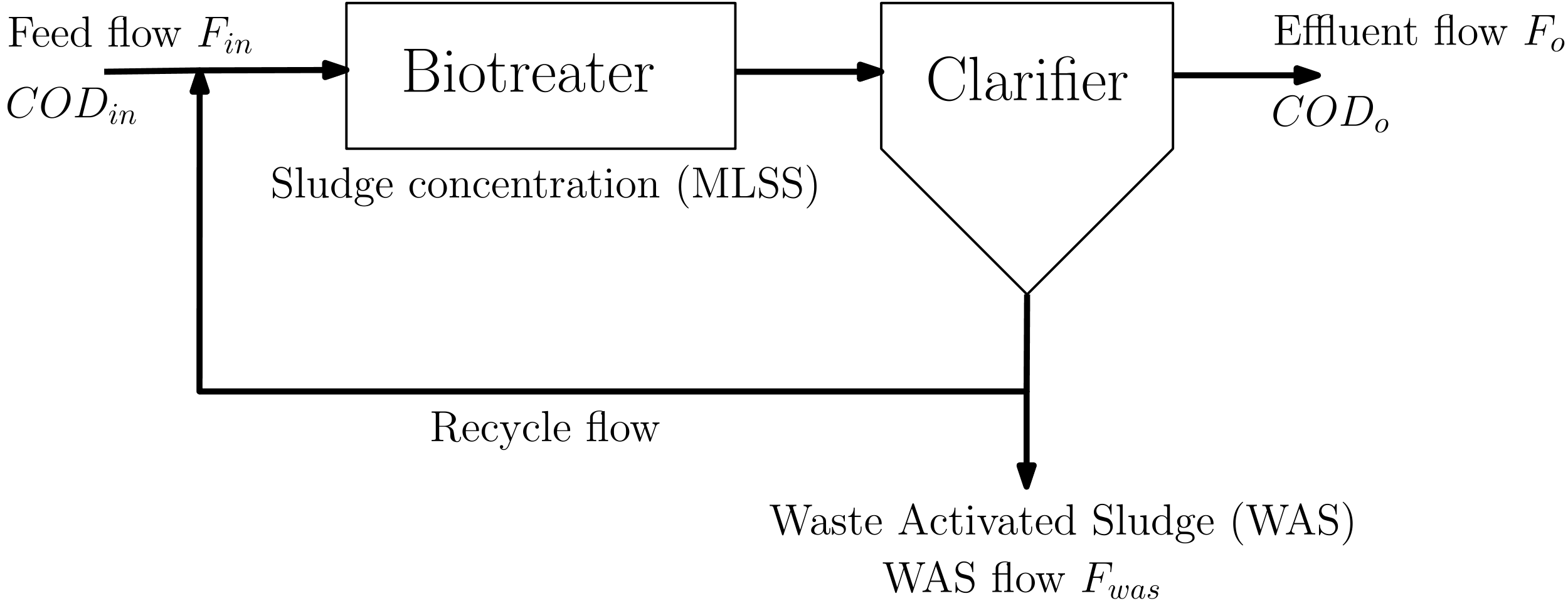
In chemical plants, the effluent streams are collected from multiple process units. Therefore, modelling and prediction of the performance of the biotreaters are strongly influenced by the uncertainties in the upstream processes (plant disturbances / shutdowns). Under such circumstances, even highly non-linear models, such as artificial neural networks (ANN), show reduced predictive performance due to sudden abnormal changes in the effluent quality (Sung et al., 2002). Other factors include reliability of process and lab measurements (Puig et al., 2008), and environmental conditions like temperature and humidity (Sayigh and Malina, 1978). The process is in a continuous non-steady state and the ability to accurately predict these input variables will directly impact the predictive capabilities of the biotreater forecasting models. Consequently, the availability of reliable forecasts and automated control strategies are expected to enhance the performance of the biotreater (Sarna et al., 2023). Though it is impossible to model unplanned shutdowns or plant disturbances in the upstream units, one can still model the progressive effect of these events. Time-series forecasting (Kotu and Deshpande, 2019) can offer a robust alternative to model the effect of environmental factors on the process.

Figure 1: Schematic diagram

The objective of this work is to develop a reliable adaptive dynamic model that can be used for forecasting the biotreater KPIs and thus, can help in making informed decisions. Since developing a physics-based model is challenging in an industrial setup (Sung et al., 2002), the aim is to develop a data-driven model which captures the correct correlations and incorporates appropriate process delays between different variables of interest. We use forecasting methods from timeseries analysis to forecast the exogenous inputs that are required to forecast the KPIs. Along with the forecasts, uncertainties associated with the predictions are also computed.

* 1. System Description

A simplified schematic of the system is given in Figure 1. The feed flow containing contaminants (mainly hydrocarbons) enter the biotreater and the biomass (measured as Mixed Liquor Suspended Solids or MLSS) inside biotreater feeds on the contaminants in the feed. The amount of contaminants in the water are measured in terms of the Chemical Oxygen Demand (COD). As the feed flows through the biotreater basins, biomass reduces the COD level and produces more biomass. The job of the clarifier is to settle the solid biomass so that pure water with low COD levels comes out of the clarifier in the effluent flow. A part of the settled biomass (sludge blanket) taken out from the bottom of the clarifier; is recycled back into the biotreater (Recycle Activated Sludge or RAS) and the remaining part is removed from the system (Waste Activated Sludge or WAS). Since COD conversion takes place in the biotreater and the chemical content is not changed in the clarifier, COD levels in the outlet of biotreater and clarifier are assumed to be the same in our analysis. The list of measured variables used for system identification are given in Table 1. Some measurements are available real-time through online sensors, while some are intermittent measurements obtained from lab analysis. In addition, the measurements of COD and MLSS tend to have high variance.

* 1. Data-driven Modelling

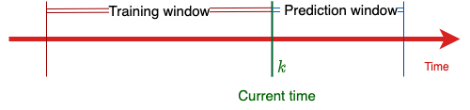
In this section, we discuss the proposed data-driven modelling framework for ETP. In addition to modelling the biotreater and clarifier operation, we also need to forecast the input variables to the system if we are to use the model for forecasting the desired output features. Additionally, quantifying the uncertainties associated with the predictions of the model and the forecasted inputs can be valuable in taking operational decisions based on the predictions.

Figure 2: Online modelling

* + 1. Dynamic Model for Biotreater & Clarifier

One of the challenges in constructing data-driven models for the biotreater system, as with many other dynamical systems is the fact that the system behavior changes with time, and the parameters estimated from historical data obtained historically may not be applicable to describe the system now. This means that a single model developed using a single set of historical data may not work well. To address this, we update the model whenever a new batch of data is made available leading to an adaptive model framework. In this paper, we assume that the process is linear for the short duration we consider in one batch of data. As the dynamics of the biotreater process are quite slow, the assumption of linearity for a short duration is reasonable. Moreover, linear models are preferred since they are simpler to handle and can be used to design efficient algorithms for control and optimization. Hence, we look at linear models which can describe the output features of both main reactor and clarifier accurately. The residence time of the biotreater is usually of the order of a few weeks, and the delays have to be properly encoded into the model for accurate predictions. We implemented Ordinary Least Squares (OLS) in an online modelling framework by using a window-based training and prediction scheme as shown in Figure 2.

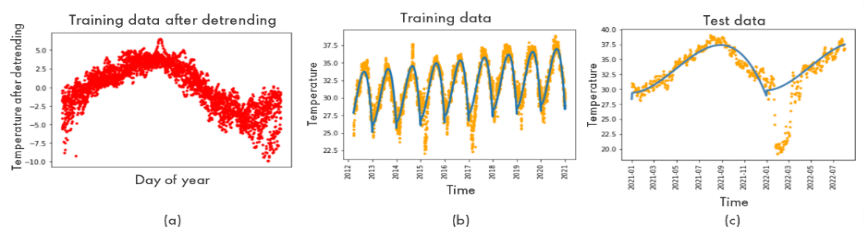


Figure 3: Temperature forecasting. (a) Model (b) Forecast on training set (c) Forecast on test set

Table 1: List of variables

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Data Source** | **Symbol** | **Units** |
| Inlet COD | Online Analyzer | CODin | ppm |
| Inlet flow | Online sensor | Fin | tons/hr |
| Biomass in biotreater | Lab analysis | MLSSb | ppmW |
| Outlet COD | Online analyzer | CODo | ppm |
| Outlet flow | Online sensor | Fo | tons/hr |
| WAS flow | Online sensor | FWAS | tons/hr |
| Sludge blanket level in clarifier | Manual | SBL | ft |
| Ambient temperature | Online sensor | Tamb | oC |

The proposed linear model is given as:

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |

where is the state vector related to the biotreater , is the state vector related to the clarifier, is the independent variable, and *vi* represent various disturbance variables. Here, . are constant matrices of appropriate sizes, are integers representing the time delays between the states and the corresponding input or disturbance variables. The above model is essentially an ARX (auto-regressive exogenous) model with fixed lags for exogenous inputs. The objective of the modelling exercise is to estimate from data. The modelling proceeds by first determining the lags using process knowledge. If we have a good estimate for the lag from process understanding, we can use these lags. If not, we follow the steps below to determine the lags:

1. From process knowledge, determine the directional sensitivity of the given disturbance w.r.t , i.e., determine the direction of change of when a positive step change in is given. Thus, if direction of change is positive, and if the direction of change is negative. Similarly, find the directional sensitivity of the input w.r.t .
2. Determine and using and where is the cross-correlation between and and is the cross-correlation between and .
3. We repeat the same procedure for the clarifier to determine and , by choosing cross-correlation of input variables with instead of .

Once the lags are determined, the parameter matrices are then determined by using ordinary least squares (OLS).

* + 1. Input Forecasting

There are four input features of interest with respect to modeling outlet COD and MLSS in the biotreater: WAS flow (), inlet COD (), inlet flow (, and temperature (. Forecasting outlet COD and MLSS requires forecasts for these input features. WAS flow is the manipulated variable that the operators would manipulate to ensure the controlled variables (MLSS and food to mass ratio) are performing as desired. Hence, in case of WAS flow, it is of our interest to see the variations in controlled variables with specific changes in WAS flow. For forecasting of output variables, we set WAS flow to be a specific constant value of interest.

Inlet flow () and inlet COD () are dependent on the upstream processes. We forecast these input variables using an auto-regressive (AR) model of order :

|  |  |
| --- | --- |
|  | (3) |

where represents the error in for sample . Temperature is modelled as a combination of linear trend and a cubic model (which models the periodic behaviour throughout the year). This model is trained using the historical data available for 8-9 years. One can also estimate the variance in temperature prediction for every day of the year from this training set. Figure 3a shows the plots for temperature forecasting and 3b shows the forecasting results. Note that the model is able to capture the seasonality as well as the increasing trend in time.

* + 1. Uncertainty Quantification

To quantify the uncertainty associated with model predictions (or forecast), we estimate the variance of each prediction. For a fixed lag ARX model of the form

|  |  |
| --- | --- |
|  | (4) |

where for all are inputs and is the error in . After re-writing as a function of and assuming that variance in is zero (since is known), the covariance matrix of can be derived as follows:

|  |  |
| --- | --- |
|  | (5) |

Assuming 95% confidence interval, uncertainty in sample prediction of variable () can be quantified as where is the standard deviation of obtained from the diagonal element in . This approach can be used to quantify uncertainty in both and . To evaluate this, we require the covariance matrix for error and all inputs (and ). The error covariance matrix () can be estimated from the training set. Since WAS flow () is a manipulated variable (value set by the user), standard deviation for the same can be set to zero for all time. Since the input variables and are modelled using AR model, their variances can be computed as explained for the ARX model.

|  |  |
| --- | --- |
|  | (6) |

|  |  |
| --- | --- |
| A graph of different weather conditions  Description automatically generated with medium confidence  (a) | A graph of different weather conditions  Description automatically generated with medium confidence (b) |

For modeling uncertainty in temperature, the variance can be estimated from the training data for each day of the year during model building process. By mapping with the corresponding day of the year, can be computed. Using the variances obtained for all input features and error, variance in can be computed and then used to obtain prediction uncertainty interval of .

Figure 4: Predictions of KPIs using (a) true inputs (b) forecasted inputs

* 1. Results

In this section, we discuss the modelling results on an industrial biotreater and compare the predictions obtained from model simulation with the actual measurements. A training window of 100 days and a model update window of 21 days are used in the simulations. The modelling parameters used in this study are given in Table 2. The model presented in this paper is obtained using OLS and the model is updated at regular intervals so that it adapts to the changing conditions in the plant. The training window refers to the number of samples used for training a model, and the update window refers to the time after which the model is updated. Thus, each model is used to predict for 14 days after which an updated model based on the more recent data is generated.

Table 2: RMSE for predictions

|  |  |  |  |
| --- | --- | --- | --- |
| RMSE ( score) | CODin | MLSSb | SBL |
| True inputs | 3.88 (0.54) | 113.75 (0.94) | 0.27 (0.79) |
| Forecasted inputs | 3.99 (0.52) | 134.65 (0.92) | 0.28 (0.78) |

We compare the results for two cases: (1) when the true input variables (inlet flow, inlet COD, ambient temperature) are used for predicting the states, (2) when the forecasted inputs are used for predicting the states. The predictions for the true input case are show in Figure 4a. As can be seen, the model starts predicting after about 100 days, the data prior to that being used for training the first model. The actual measured values are also plotted along with the predicted values for comparison. Once the model is built, it will be used for prediction for the next 21 days. Then a new model is built and used for prediction for the next 21 days. The total time of predictions shown in the plot is 40 days. The predictions for the forecasted input case are shown in Figure 4b. The corresponding forecasts for the inputs are given in Figure 5. The forecasts try to capture the mean trend in the variables based on the immediate past. This is effective if the upstream processes are in a steady state. We can also see a jump in the predictions after 21 days due to model update. The Root Mean Squared Error (RMSE) and scores for the predictions are shown in Table 2.

|  |  |
| --- | --- |
|  |  |
| Figure 5: Forecast of input variables | |

* 1. Conclusion

In conclusion, this work aims at building data-driven dynamical models for an industrial biotreater and clarifier system. The model is expected to be used for forecasting the states and for model predictive control. For forecasting KPIs, forecasts of exogenous inputs are used. The proposed models can capture the correct correlations between inputs and outputs. Future work incudes improvement of the model by attempting recursive linear models or nonlinear models. Implementation of control algorithm for ensuring required food to mass ratio is part of the future work.

References

Sayigh, B.A., Malina, J.F., 1978. Temperature effects on the activated sludge process. Journal (Water Pollution Control Federation), 678-687.

Eckenfelder, W., Musterman, J., 1998. Activated sludge: Treatment of industrial wastewater. Technology and Engineering, CRC Press.

Ioannou, P., Fidan, B., 2006. Adaptive Control Tutorial. Advances in Design and Control, Society for Industrial and Applied Mathematics.

Kotu, V., Deshpande, B., 2019. Time series forecasting. Data Science.

Newhart, K.B., Holloway, R.W., Hering, A.S., Cath, T.Y., 2019. Data-driven performance analyses of wastewater treatment plants: A review. Water Research, 157, 498–513.

Puig, S., M.C.M, V.L., J, C., S.C.F, M., 2008. Data evaluation of full-scale wastewater treatment plants by mass balance. Water research, 42, 4645–4655.

Sarna, S., Patel, N., Corbett, B., McCready, C., Mhaskar, P., 2023. Process-aware data-driven modelling and model predictive control of bioreactor for the production of monoclonal antibodies. The Canadian Journal of Chemical Engineering, 101, 2677–2692.

Sung, L.D., Jeon, C.O., Park, J.M., Chang, K.S., 2002. Hybrid neural network modelling of a full-scale industrial wastewater treatment process, Biotechnology and Bioengineering, 78, 670-682.

Sunita, V., Joshi, R., Srivastava, V.K., Ngo, H.H., Guo, W., 2020. Treatment of wastewater from petroleum industry: current practices and perspectives, Environmental Science and Pollution Research, 27, 27172-27180.