Design of a hybrid mechanistic/Gaussian process model to predict full-scale wastewater treatment plant effluent

Nadja Hvala[a,*], Juš Kocijan[a,b]

[a] Jožef Stefan Institute, Department of Systems and Control, Jamova cesta 39, SI-1000 Ljubljana, Slovenia

[b] University of Nova Gorica, Vipavska cesta 13, SI-5000 Nova Gorica, Slovenia

Abstract

This paper presents the design of a hybrid model of a wastewater treatment plant (WWTP), which is meant to improve the quality of effluent prediction. By combining mechanistic, i.e. activated sludge model, and data-driven model it is expected to retain physical transparency and achieve good prediction accuracy. For the data-driven model, a state-of-the-art machine learning approach based on Gaussian process (GP) model was applied. GP models systematically address model uncertainty when lacking identification data and are applicable also for small data-sets, which both are encountered in WWTP modelling. Serial and parallel hybrid structures were designed to address the challenges of missing input data, insufficient mechanistic model accuracy and demanding model parameter estimation. Results of full-scale effluent predictions show that, by applying hybrid models, the accuracy of the model is improved. Good results were obtained also for default values of activated sludge model parameters, which significantly simplifies the model design process.

Keywords:

Wastewater treatment; Hybrid model; Gaussian Process model; Machine learning; Model prediction

[*] Corresponding author.
E-mail address: nadja.hvala@ijs.si (N. Hvala)

1. Introduction

Mathematical modelling is one of the most fundamental scientific disciplines that is used for better understanding, analysing and predicting the performance of different systems and processes. Mathematical modelling of biological wastewater treatment processes has a long tradition and is indispensable in the design and operation of wastewater treatment plants (WWTPs) for testing different operation scenarios and optimising the plant performance. When applying the models in control and decision-making, the model accuracy concerning the intended model use is one of the key questions to consider.

Generally, two modelling approaches are possible, i.e. a theoretical approach or a data-driven approach. In the case of wastewater treatment, mechanistic models, also called first-principle or theoretical models are commonly used. The state-of-the-art models are a family of activated sludge models (ASM) (Henze et al., 2000), which conceptualize theoretical knowledge of the biological processes and other observed phenomena occurring within the system. ASM models are capable of extrapolating the process performance in a wide variety of process operating conditions. They also give a deeper insight into the process since many process variables are not directly measured or observed on-line. Data-driven modelling, on the other hand, is more generally applicable to different processes and well supported by efficient machine learning techniques. It may also result in a more accurate model performance if any limitations arise from a pre-set structure of a first-principle model, which as such could not capture all information latent in data. An overview of data-driven techniques applied in WWTPs indicates that the prevailed methods used are multivariate statistics, artificial neural networks (ANNs) and fuzzy systems (Corominas et al., 2018). Data-driven models are mainly used for prediction and soft-sensing, process monitoring and fault detection.

Both modelling approaches have disadvantages. The usage of a mechanistic model in full-scale WWTP applications requires extensive and time-consuming adjustment of model parameters to real plant behaviour where significant computational effort and expert knowledge may be needed for model parameterisation. Besides, parameter estimation based on numerical optimization algorithms often gives non-unique parameter estimates and many local optima because of the ill-conditioned, non-convex optimization problem (Manheim and Detwiler, 2019). Data-driven modelling has other difficulties. Data-driven models are usually designed for a limited sub-set of process variables and a limited sub-set of process operating conditions. Besides, they are not able to describe the process physically transparently.

To take the strength of both modelling approaches and overcome their deficiencies, hybridisation combining first-principle and data-driven models is proposed (Anderson et al., 2000). The hybrid model is expected to retain the first-principle model structure with all the modelled process variables. Besides, it should retain physical transparency and relationships between the variables. The hybrid model is also expected to enhance the prediction accuracy of the selected observed process variable, or several variables, for which a more accurate prediction is desired. Combinations of first-principle and data-driven models are particularly enhanced with increased monitoring and sensors implementations, which enable the design of data-driven models (Newhart et al. 2019). The trend of increased sensors development and implementation is also evident in wastewater treatment (Kruse, 2018).

Applications of hybrid models in wastewater treatment are still quite rare (Haimi et al., 2013). In particular, very few research papers address hybrid models by combining mechanistic ASM models and data-driven models. Côté et al. (1995) designed a feedforward neural network model that simulated the prediction errors of a simple, previously tuned mechanistic model. Anderson et al. (2000) presented two hybrid models where feedforward neural networks were used to predict the process reaction rates and correct the prediction error of the linearized ASM1 model. An interesting remark is that the application of their approach to a more complex problem involving phosphorus kinetics was not successful. Lee et al. (2002) also designed a hybrid neural network based on a simplified mechanistic model and presented good extrapolation properties of the hybrid model.

As presented above, the hybrid approaches applied so far were based on artificial neural networks for the data-driven model in a hybrid structure. In this paper, the data-driven part of the hybrid model is based on Gaussian Process (GP) model. GP modelling is a machine learning probabilistic modelling approach with a specific property that the mapping between regression inputs and output is presented with a stochastic process and, consequently, the model prediction with a distribution, expressed in terms of mean value and the variance of the modelled variable. The variance indicates the model prediction uncertainty and is used as a confidence measure for the model prediction results. The reasons for using GP models for the construction of a data-driven part of the hybrid model are many (Kocijan, 2016). Firstly, GP models, as kernel models, contain noticeably fewer parameters to optimise than frequently used statistic models based on basis functions. Secondly, they are very convenient for parameter optimisation in cases where the number of data points is relatively small. Thirdly, GP models have comparable prediction capabilities to ANNs but account for model uncertainty (Bradford et al., 2018). For example, by indicating the higher variance around the predicted mean, the GP model can highlight areas of the input space where model-prediction quality is poor, e.g. due to the lack of training data or its complexity. In this way, the GP model can quantify the model quality systematically. These properties of GP models are important in wastewater treatment modelling. Model tuning and optimization are most often performed on a limited set of data that is constrained by the number of laboratory measurements for some variables. Besides, WWTP process input and operating parameters change with time and may vary in a wide region that was not incorporated within the identification data. The application of GP models in wastewater treatment is still rare. Most often GP models are used for monitoring and fault detection, e.g. sensor drift detection (Samuelsson et al., 2017), secondary settler monitoring (Zambrano et al., 2019), filamentous sludge bulking prediction (Liu et al., 2016), prediction of biodegradation completion time (Kocijan and Hvala, 2013), or membrane fouling (Chan et al., 2015).

This paper aims to present the application of GP models in hybrid modelling of WWTPs, which has not been considered yet. The case study is the design of process models to predict effluent total nitrogen and total phosphorus concentrations in a full-scale WWTP. Model-based prediction of effluent concentrations is of interest for different purposes. It is used in predictive control schemes for energy consumption and performance optimisation (Vrečko et al., 2011). The prediction of effluent is used to detect a risk of violation of effluent limits enough time in advance to select a suitable control strategy (Santín et al., 2015). It can be also used as an early warning prediction tool for water quality control in combined waste and wastewater treatment processes (Guo et al., 2015). The main contribution of our work is an attempt to improve the mechantistic model prediction by the support of the GP model. The usage of mechanistic model is preferred over a completely data-driven model since it incorporates knowledge of the process fundamentals (Anderson et al., 2000). Hence, the aim is not to obtain the best model per see, but to refine the prediction of ASM model with the additional information potentially present in the collected process data that is otherwise not captured by the mechanistic model. In this paper, GP models are used in serial and parallel hybrid structures to complement mechanistic models. The data-driven model supports the mechanistic model in two commonly encountered limitations of full-scale WWTP modelling studies, i.e. the problems of missing input data and demanding model parameter estimation for model-data calibration. As also reported in the literature, scarce data sets measured at the inlet of WWTP (Martin and Vanrolleghem, 2014) and the uncertainty of model parameters (Mannina et al., 2012) are among the main limitations for more widespread utilization of WWTP models.

2. Methods

2.1. Data-driven models

Data-driven models are used for modelling of dynamic systems as the alternative to first-principle models also in the case of WWTPs (Haimi et al., 2013, Corominas et al., 2018, Newhart et al., 2019). Among methods for data-driven modelling predominantly the methods are found, where the system is approximated by a linear or nonlinear combination of some *basis functions* with coefficients that have to be estimated. Examples of such data-driven models are artificial neural networks, fuzzy models, Volterra-series models, etc. Depending on the nonlinearity, a fixed basis function approach could need a relatively large number of basis functions to approximate the unknown nonlinear system. The increase in the number of basis functions and consequently the increase of coefficients to be estimated requires also a large number of data necessary for coefficients estimation. Data overfitting might occur when the number of data is not appropriate (Manheim and Detwiler, 2019).

The alternative to data-driven models with basis functions that circumvents the mentioned disadvantage is *kernel methods*. Kernel methods (Bishop, 2006) do not try to approximate the modelled system by fitting the parameters of the selected basis functions, but rather they search for the relationship among the measured data. The model is composed of input-output data that characterises the behaviour of the modelled system and the kernel function that describes the relation of the output data concerning the input data.

Data-driven models depend on observation data used for modelling. These data may be noisy and affected by various disturbances. Probabilistic modelling (Peterka, 1981, Green et al., 2015), namely Bayesian modelling, is a way to treat uncertainties, model complexity and reduce the level of overfitting.

2.2. Gaussian Process models

GP models are probabilistic, data-driven, kernel models based on the principles of Bayesian probability. The following additive system is considered when GP modelling (Rasmussen and Williams, 2006, Shi and Choi, 2011, Kocijan, 2016) is used for regression:

$$y = f(\mathbf{z}) + \nu, \qquad (1)$$

where y is process output, ν is white Gaussian noise and $\mathbf{z}$ is the vector of regressors from the operating space $\mathbb{R}^D$, D is the number of regressors. The noise is of the form $\nu \sim \mathcal{N}(0, \sigma_n^2)$ where σ_n^2 is the variance. Elements of the vector $\mathbf{z} \in \mathbb{R}^D$, i.e. $z_i: i = 1, \dots, D$ are called *regressors* and the vector $\mathbf{z}$ is called the *regression vector*.

We look for a nonparametric Bayesian model, where the function to be modelled f is set as a GP

$$f(\mathbf{z}) \sim \mathrm{GP}\left(E(f(\mathbf{z})), \mathrm{cov}\left(f(\mathbf{z}_i), f(\mathbf{z}_j)\right)\right), \tag{2}$$

where $E(f(\mathbf{z}))$ is a mean of function f and $\mathrm{cov}\left(f(\mathbf{z}_i), f(\mathbf{z}_j)\right)$ is the covariance of function f.

Mean and covariance define the properties of the stochastic process that we model. They incorporate the prior knowledge of stochastic process to the system training. For the sake of simplicity, we often assume the mean function is selected as 0. The covariance matrix is calculated using *covariance functions*, i.e., kernel functions, which are characterised with hyperparameters.

An overview of some of the possible covariance functions is given in (Rasmussen and Williams, 2006). The covariance functions $C(\mathbf{z}_i, \mathbf{z}_j)$ that we use in this paper are the following:

- Linear covariance function (3) with ARD option (Kocijan, 2016)

$$C(\mathbf{z}_i, \mathbf{z}_j) = \mathbf{z}_i \Lambda_{LIN}^{-1} \mathbf{z}_j, \tag{3}$$

where Λ_{LIN}^{-1} is a diagonal, semi-definite matrix with weights that implements the use of different length scales on different regressors and can be used to assess the relative importance of the contributions made by each regressor. This is a property called *Automatic Relevance Determination* (ARD) (Rasmussen and Williams, 2006).

- Exponential covariance function (4) with ARD option

$$C(\mathbf{z}_i, \mathbf{z}_j) = \sigma_f^2 e^{\left[-\frac{1}{2}(\mathbf{z}_i - \mathbf{z}_j)^T \Lambda_{SE}^{-1} (\mathbf{z}_i - \mathbf{z}_j)\right]} \tag{4}$$

where Λ_{SE}^{-1} is a diagonal, semi-definite matrix, while the hyperparameter σ_f^2 represents the scaling factor of the possible variations of the function.

- Matérn covariance function (5) with ARD option

$$C(\mathbf{z}_i,\mathbf{z}_j) = \sigma_f^2\left(\frac{2^{1-d}}{\Gamma(d)}\right)\left(\frac{\sqrt{2d}r}{l}\right)^d K_d\left(\frac{\sqrt{2d}r}{l}\right), \tag{5}$$

where Γ is the gamma function, the hyperparameter l or the horizontal scaling factor determines the relative weight on distance for the input variable $\mathbf{z}$, K_d is a modified Bessel function, the hyperparameter d controls the differentiability of the modelled mapping function and r is a distance function

$$r = \sqrt{(\mathbf{z}_i - \mathbf{z}_j)^T \Lambda_M^{-1}(\mathbf{z}_i - \mathbf{z}_j)}, \tag{6}$$

where $\Lambda_M^{-1}$is a diagonal, semi-definite matrix.

A covariance matrix $\mathbf{K}$ is calculated by evaluating the covariance function given all the pairs of measured data. The elements K_{ij} of the covariance matrix $\mathbf{K}$ are covariances between the values of the functions $f(\mathbf{z}_i)$ and $f(\mathbf{z}_j)$ corresponding to the arguments $\mathbf{z}_i$ and $\mathbf{z}_j$,

$$K_{ij} = \operatorname{cov}\left(f(\mathbf{z}_i), f(\mathbf{z}_j)\right) = C(\mathbf{z}_i,\mathbf{z}_j). \tag{7}$$

This means that the covariance between the random variables that represent the outputs, i.e., the functions of the arguments, numbers i and j, equals the covariance function C between the arguments, numbers i and j.

The data for the training of the model is described as a data set $\mathcal{D} = \{(\mathbf{z}_i, y_i)|i = 1, \dots, N\} = \{(\mathbf{Z}, \mathbf{y})\}$, where *N* is the number of observations. Following the Bayesian modelling framework, we are looking for the posterior distribution over f, which for the given data $\{(\mathbf{Z}, \mathbf{y})\}$ and hyperparameters $\boldsymbol{\theta}$ of the used covariance function is

$$p(f|\mathbf{Z}, \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|f, \mathbf{Z}, \boldsymbol{\theta})p(f|\boldsymbol{\theta})}{p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\theta})}, \tag{8}$$

where $p(\mathbf{y}|f, \mathbf{Z}, \boldsymbol{\theta})$ is the likelihood, $p(f|\boldsymbol{\theta})$ is the prior probability distribution of function f for the given hyperparameters $\boldsymbol{\theta}$, $p(\mathbf{y}|\mathbf{Z}, \boldsymbol{\theta})$ is the evidence or marginal likelihood and $p(f|\mathbf{Z}, \mathbf{y}, \boldsymbol{\theta})$ is the posterior distribution over f.

The Bayesian inference (8) of most systems can only be implemented using analytical or numerical approximation. One possible approximation method is the estimation of hyperparameters with the maximisation of the evidence. See (Rasmussen and Williams, 2006, Kocijan, 2016) for details.

The objective of the modelling is to find the predictive distribution of the latent function values $f^*= f(\mathbf{z}^*)$ at test inputs $\mathbf{z}^*$. The posterior predictive distribution of f^* is obtained by marginalising the function out. The resulting predictive distribution is Gaussian and defined with equation

$$p(f^*|y) \;=\; \mathcal{N}\,(\mathbf{K}_*(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\,\mathbf{y}, \mathbf{K}_{**} - \mathbf{K}_*(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\mathbf{K}_*), \tag{9}$$

where $\mathbf{I}$ is the unity matrix. Therefore, the mean $E(f^*)$ and variance $\mathrm{var}(f^*)$ of predictive distribution at $\mathbf{z}^*$ are given as

$$E(f^*) \;=\; \mathbf{K}_*(\mathbf{K} + \sigma_n^2\mathbf{I})^{-1}\,\mathbf{y}, \tag{10}$$

$$\mathrm{var}(f^*) \;= \mathbf{K}_{**} - \mathbf{K}_*(\mathbf{K} + \sigma_n^2\mathbf{I})^{\mathbf{-1}}\mathbf{K}_*, \tag{11}$$

where $\mathbf{K}_* = [C(\mathbf{z}_1, \mathbf{z}^*), \dots, C(\mathbf{z}_N, \mathbf{z}^*)]^{\mathrm{T}}$ is the $N \times 1$ vector of covariances between the training input data and the test input data, and $\mathbf{K}_{**} = C(\mathbf{z}^*, \mathbf{z}^*)$ is the autocovariance of the test input data. The variance $\mathrm{var}(f^*)$ is composed of the variance due to measurement noise and the variance due to lack of data for modelling.

2.3. Hybrid models

Hybrid models are achieved by combining different models. In this paper, hybrid models refer to combined theoretical, i.e. mechanistic models and data-driven models.

Hybrid modelling can be performed based on different hybrid structures, serial, parallel, or parallel-serial one (Hajirahimi and Khashei, 2019). The serial configuration uses data-driven models to complement poorly defined parts of the first-principle models, in most cases kinetic terms of mechanistic models. The parallel configuration uses the data-driven model to model the residuals from the first principles (Anderson et al., 2000). Fig. 1 shows the serial and parallel hybrid structures with mechanistic models and GP models. It should be emphasised that other data-driven models can also be used.

The serial hybrid structure was in our case used for data reconciliation and gap-filling of WWTP input data. The knowledge of WWTP input, i.e. input wastewater composition and concentrations, is crucial for the prediction of WWTP effluent concentrations by the mechanistic model. Since input data rely on laboratory and on-line measurements, they are often erroneous or lacking. Different pre-treatment approaches can be used to solve the problem of missing or outlier data, e.g. data tagging and replacing with data filling such as interpolation, correlation with another measured value, daily average, the day before data, influent model, etc. A recent evaluation of the different data analysis and gap-filling methods shows that using the influent model to fill gaps in the data yields the highest reliability (Mulder et al., 2018). An example of estimating missing data in a WWTP flow rate signal by the GP regression model has been also successfully tested on simulated data (Samuelsson et al., 2017). In this paper, the data-driven GP model in the serial hybrid structure is used to complement influent data, if the laboratory or on-line measurements of input wastewater are erroneous or not available.

The parallel hybrid structure was used to improve the model prediction accuracy. In this case, the data-driven part of the hybrid model is used to estimate and thus compensate for the mechanistic model residual, i.e. the error between the mechanistic model and the measured process output.

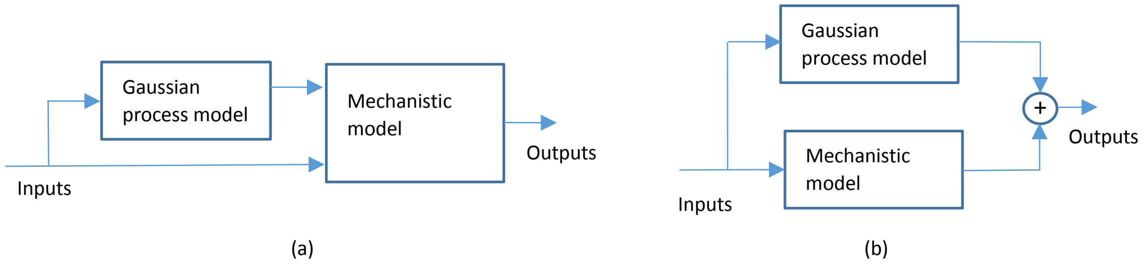

Fig. 1. Hybrid modelling structures combining the mechanistic model and Gaussian process model: (a) serial and (b) parallel configuration.

3. WWTP case study

The plant considered in this study is a large full-scale WWTP operated at 435,000 population equivalent (PE). The treatment facilities consist of mechanical treatment (screens, grit and grease chamber), a biological stage with suspended biomass activated sludge process (three parallel plug-flow aerobic reactors and four parallel secondary clarifiers) and sludge treatment (sludge thickening, anaerobic digestion, dewatering and sludge drying).

The biological stage of the plant removes carbon and achieves nitrification, and will be upgraded for complete nitrogen and phosphorus removal. For plant upgrading, the mechanistic model of the plant was designed and tuned to plant measurements (Hvala et al., 2018). The model can be also used in the on-line operation to predict effluent concentrations. The mechanistic model of the given case study is considered sufficiently reliable for the evaluation of plant performance in different operating scenarios, but in predicting effluent concentrations some major differences between the model and the process are still present.

Fig. 2 shows the process configuration of the biological stage and the available process data for modelling. As presented in the next sections, serial and parallel hybrid structures based on the mechanistic model and different GP models are designed to predict effluent total nitrogen (*TN*) and total phosphorus concentrations (*TP*). For each particular model structure, a subset of process input variables is used.

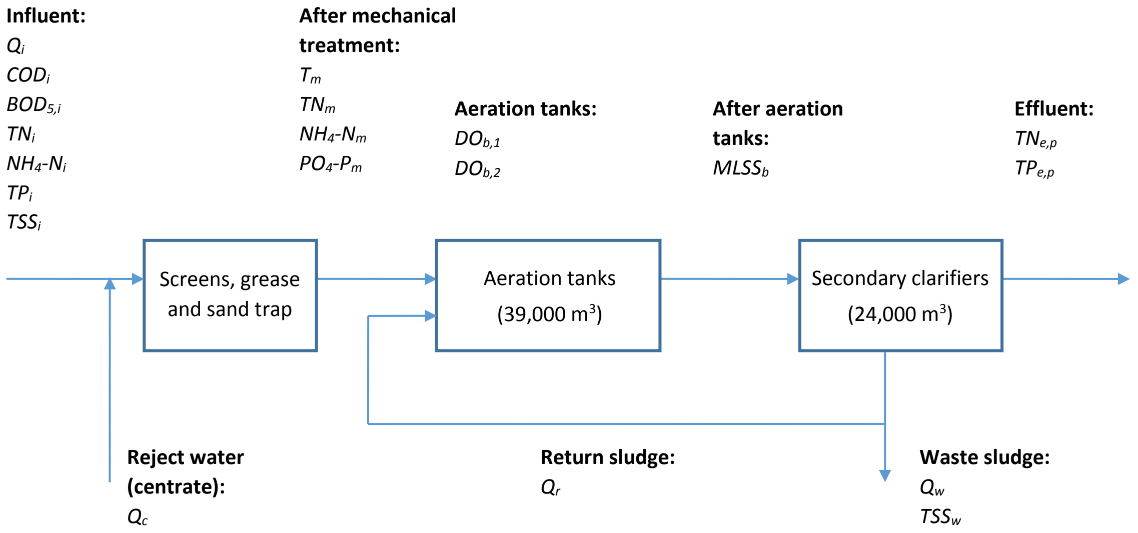

Fig. 2. The full-scale WWTP biological stage with the indicated flow (Q) and concentrations measurements that are used for modelling: chemical oxygen demand (COD), biochemical oxygen demand in five days (BOD_5), total nitrogen (TN), ammonia nitrogen (NH_4-N), total phosphorus (TP), total suspended solids (TSS), wastewater temperature (T), orthophosphate (PO_4-P), dissolved oxygen (DO) and mixed liquor suspended solids (MLSS).

3.1. Mechanistic model

The mechanistic model is designed as a plant-wide model and is already presented in (Hvala et al., 2018). The aerobic reactors in the biological stage are modelled with ASM2d model, while secondary clarifiers are considered as biologically inactive and are modelled with the double exponential settling velocity function (Takács et al., 1991). The model is designed in GPS-X simulation software (Hydromantis, 2016).

The list of input and output variables of the mechanistic model is shown in Table 1. The model output variables are effluent total nitrogen ($TN_{e,m}$) and total phosphorus ($TP_{e,m}$) concentrations and are compared to process measured values ($TN_{e,p}$ and $TP_{e,p}$). As can be seen, not all the available process measurements presented in Fig. 2 are directly used as inputs of the mechanistic model. For example, influent BOD_5 and *TSS* measurements are not used, since the input wastewater is characterized by the *COD* model, therefore, these two variables are the outputs of the COD model. Similarly, also the N and P measurements after the mechanical treatment are not used since they are the outputs of the model. The model was tuned to plant measurements by adjusting nine wastewater composition parameters and three kinetic and stoichiometric parameters of the model.

Table 1. Mechanistic model input and output variables. Index k refers to discrete-time with time step one day.

Mechanistic model inputs	**Mechanistic model outputs**
$Q_i(k)$	$TN_{e,m}(k)$
$COD_i(k)$	$TP_{e,m}(k)$
$TN_i(k)$	
$NH_4\text{-}N_i(k)$	
$TP_i(k)$	
$T(k)$	
$DO_{b,1}(k)$	
$DO_{b,2}(k)$	
$MLSS_b(k)$	
$Q_r(k)$	

3.2. The serial hybrid GP model

A set of data-driven GP models in the serial hybrid structure was designed to predict influent *COD*, *TN*, $NH_4\text{-}N$ and *TP* concentrations in the gaps of missing measurements. The GP model was constructed for each influent variable separately.

The estimated value of influent concentration is determined from the predicted flow and past concentration and flow measurements. For each variable, the three delayed measurements were considered as potential regressors. Regressors were selected with a sequential forward selection method as an example of a wrapping method (May et al., 2011) with 4-fold cross-validation on the training dataset. The performance measure used was the squared error of predictions regarding the measurements.

The covariance function used in serial GP models was the sum of a squared exponential (4) and a linear covariance functions (3) both with ARD option. This covariance function was selected because it provided the best results on identification dataset. The covariance function is expressed as

$$C(\mathbf{z_i},\mathbf{z_j}) = \sigma_f^2 (e^{\left[-\frac{1}{2}(\mathbf{z}_i-\mathbf{z}_j)^T \Lambda_{SE}^{-1}(\mathbf{z}_i-\mathbf{z}_j)\right]} + \mathbf{z}_i \Lambda_{LIN}^{-1} \mathbf{z}_j). \tag{12}$$

The finally selected regressors for the data-driven GP models in the serial hybrid structure are shown in Table 2. The selection of regressors and GP hyperparameters was performed on data subsets within the total dataset, where measurements of particular process input variables were available.

Table 2. Input and output variables of data-driven GP models when used for estimating influent COD, TN, NH_4-N and TP concentrations in a serial hybrid structure.

GP model inputs	**GP model output**
$\mathbf{z} = \begin{bmatrix} COD_i(k-1) \\ Q_i(k) \\ Q_i(k-1) \end{bmatrix}$	$y = COD_i(k)$
$\mathbf{z} = \begin{bmatrix} TN_i(k-1) \\ Q_i(k) \\ Q_i(k-1) \end{bmatrix}$	$y = TN_i(k)$
$\mathbf{z} = \begin{bmatrix} NH_4{-}N_i(k-1) \\ Q_i(k) \\ Q_i(k-1) \end{bmatrix}$	$y = NH_4{-}N_i(k)$
$\mathbf{z} = \begin{bmatrix} TP_i(k-1) \\ Q_i(k) \end{bmatrix}$	$y = TP_i(k)$

3.3. The parallel hybrid GP model

The output of the parallel GP model is the mechanistic model residual, i.e. the error between the mechanistic model and the measured process output. For each mechanistic model output, a separate parallel GP model was designed.

For selecting the model regressors, all measured variables presented in Fig. 2 and their three delayed values were evaluated as potential regressors. The effluent concentrations predicted by the mechanistic model ($TN_{e,m}$, $TP_{e,m}$) were included in the model structure by default. Again, regressors were selected with a sequential forward selection method as an example of a wrapping method (May et al., 2011) with 4-fold cross-validation on the identification dataset. The performance measure used for the regressors selection was the squared error of predictions regarding the measurements.

The covariance function used in parallel GP models was the sum of a Matérn covariance function (5) with hyperparameter $\mathrm{d} = 3/2$ (Kocijan, 2016) and a linear covariance function (3) both with ARD option. This covariance function was selected because it provided the best results on identification dataset. The covariance function is expressed as

$$C(\mathbf{z}_i, \mathbf{z}_j) = \sigma_f^2 (\left(\frac{2^{1-d}}{\Gamma(d)}\right) \left(\frac{\sqrt{2d}r}{l}\right)^d K_d \left(\frac{\sqrt{2d}r}{l}\right) + \mathbf{z}_i \Lambda_{LIN}^{-1} \mathbf{z}_j). \tag{13}$$

The finally selected model regressors for GP models in the parallel hybrid structure are shown in Table 3. For the selection of regressors and model identification, the available dataset was divided into a training dataset, also called identification dataset in system theory, and test dataset also called validation dataset in system theory.

It can be seen that influent COD_i, TN_i, NH_4-N_i and TP_i concentrations, which are the inputs of the mechanistic model, are not included in the parallel model. This indicates that these process variables are represented in the mechanistic model sufficiently well and do not bring significant additional information for the data-driven model. An exception is influent TN_i, which is used in the phosphorus model. This may be explained by the fact that the soluble part of influent TP_i, which is used in the mechanistic model for effluent prediction, was not measured but computed from the TP_i measurements using a fixed PO_4-P_i/TP_i ratio equal to 32% (Hvala et al., 2018). Hence, additional information for effluent TP_e prediction may be gained from TN_i signal.

The GP model regressors also do not include operating parameters, such as temperature T and dissolved oxygen concentrations $DO_{b,1}$ and $DO_{b,2}$, which are already included in the mechanistic model.

The GP model regressors include TSS_i, $MLSS_b$, *recycle flow* Q_r and waste sludge mass flow $\Phi_w=Q_w\cdot TSS_w$. All these variables are related to biomass and sludge concentrations, which may not be presented in the model sufficiently well.

The input regressor of the nitrogen model is also ammonia concentration after the mechanical treatment NH_4-N_m. It differs from influent NH_4-N_i because of recycling water (centrate) addition. In the mechanistic model, the centrate addition may not be well presented, since the centrate flow Q_c was not measured but estimated from the periods of centrifuge operation.

Table 3. Input and output variables of GP models for estimating mechanistic model residuals $R_{TN,e}$ and $R_{TP,e}$ in a parallel hybrid structure.

GP model inputs	**GP model output**
$\mathbf{z}=\begin{bmatrix} TN_{e,p}(k-1) \\ MLSS_b(k-3) \\ Q_r(k-2) \\ Q_i(k-1) \\ TSS_i(k-1) \\ NH_4-N_m(k-1) \\ \Phi_w(k-1) \\ TN_{e,m}(k) \end{bmatrix}$	$y=R_{TN,e}(k)$
$\mathbf{z}=\begin{bmatrix} TN_i(k-3) \\ TSS_i(k-3) \\ MLSS_b(k-3) \\ Q_r(k-2) \\ MLSS_b(k-1) \\ \Phi_w(k-1) \\ TP_{e,m}(k) \end{bmatrix}$	$y=R_{TP,e}(k)$

4. Results and discussion

The presented models were trained and tested on full-scale WWTP data. The total dataset included 650 days of plant operation. For all the process variables presented in Fig. 2, daily average values were available. The serial models predicted data in the gaps of missing measurements. The parallel models predicted the residuals between mechanistic model and measurements. The covariance functions applied in serial (12) and parallel (13) GP models required to estimate 2*D*+1 hyperparameters, where *D* is the number of regressors in each model structure. The training dataset for the parallel model had 400 data points and the rest data points were used for the parallel-model test. Testing of effluent *TN* and *TP* prediction by different models was performed on 250 days of plant operation.

4.1. Evaluation criteria

The model quality was evaluated with the statistical coefficient of determination R^2 computed from the sum of squares of residuals $\mathrm{SS}_{\mathrm{res}}$ and the total sum of squares $\mathrm{SS}_{\mathrm{tot}}$ concerning the mean of the observed data $\bar{y}$

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SS}_{\mathrm{res}}}{\mathrm{SS}_{\mathrm{tot}}} = 1 - \frac{\sum_i \left(y_{m,i} - y_{s,i}\right)^2}{\sum_i \left(y_{m,i} - \bar{y}\right)^2}, \tag{14}$$

where $y_{m,i}$ represent measured values and $y_{s,i}$ simulated values of given model output.

An evaluation criterion was also a standard deviation σ of the error between model predictions and the mean of the observed data $\bar{y}$.

4.2. The hybrid model with a tuned mechanistic model

The main question considered in the study was whether the mechanistic model prediction can be improved with the hybrid model structure. For predicting effluent *TN* and *TP* concentrations three model options were tested:

a) mechanistic model prediction,

b) serial hybrid model prediction,

c) combined parallel-serial hybrid model prediction.

In the case of mechanistic model prediction, the missing and erroneous influent data was replaced by the data from the previous day.

Fig. 3 shows the prediction of effluent *TN* in the case when using a mechanistic model and a combined parallel-serial hybrid model structure. It can be seen that the hybrid model improves the performance and achieves output-prediction error closer to process measured values. The output-prediction error hereafter means the difference between the mean value of prediction and the test data. The hybrid-model output-prediction error is mostly within $\pm 2\sigma$, which indicates the quality of model prediction. Fig. 4 also shows the distribution of the output-prediction-error signal for both models. It can be seen that the variance of the prediction error in the case of the hybrid model is noticeably smaller compared to the mechanistic model.

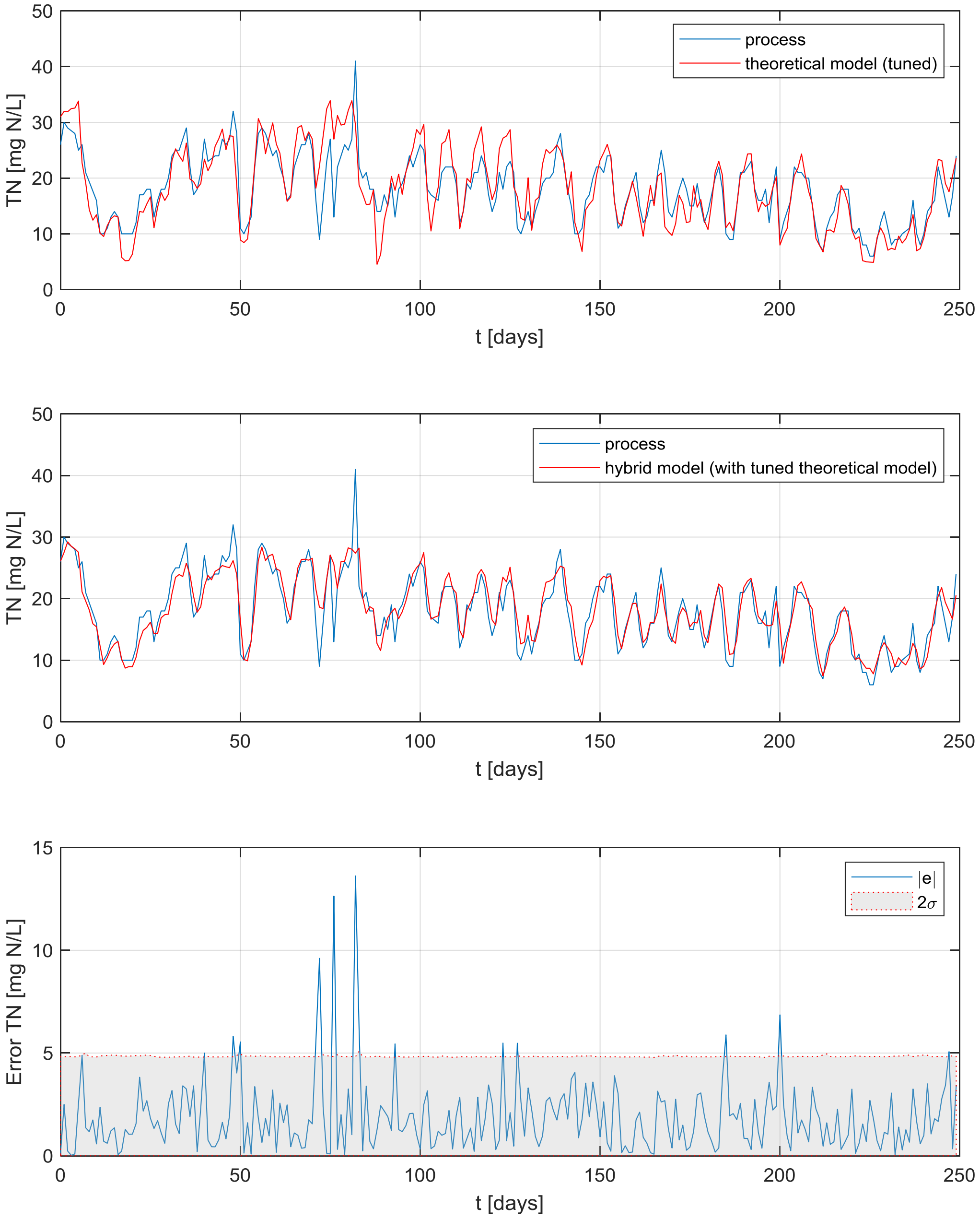

Fig. 3. Effluent total nitrogen (TN) prediction performed by the tuned mechanistic model (upper plot) and combined parallel-serial hybrid model (middle plot). The lower plot shows the absolute values of the error between the hybrid-model prediction and the true value, and 2 standard deviations of the data-driven model prediction (95 % confidence band).

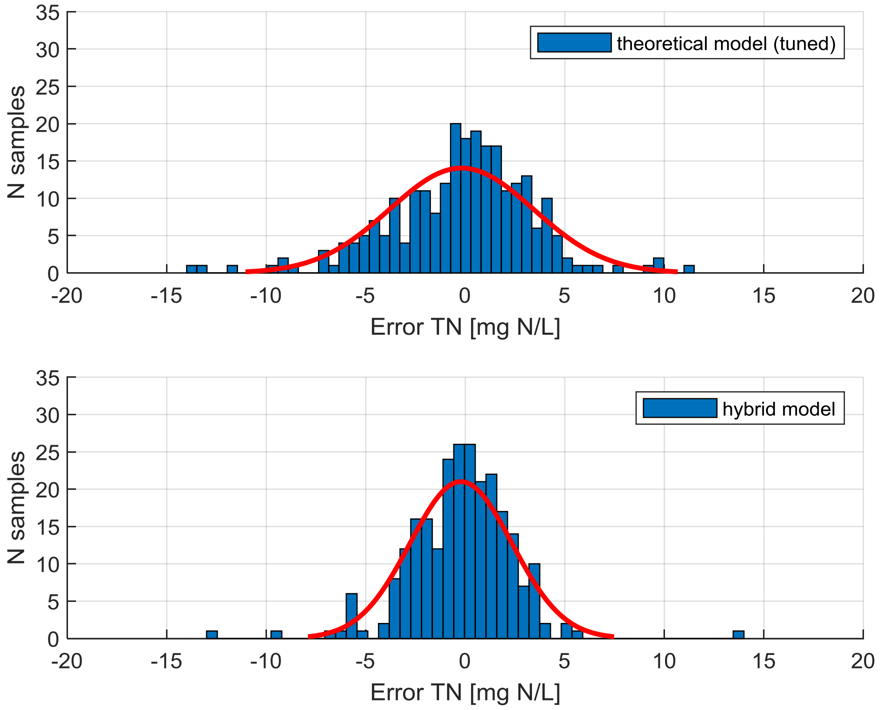

Fig. 4. Distribution of effluent total nitrogen (TN) model error in the case of the tuned mechanistic model and hybrid parallel-serial model structure based on the tuned mechanistic model.

4.3. The hybrid model with default mechanistic model

Since tuning of the mechanistic model parameters and adjusting the influent wastewater composition to a particular case-study is one of the most demanding steps of the full-scale mechanistic model design, the application of a hybrid model based on default mechanistic model was also considered. This means that in this case no adjustments of the model parameters and wastewater composition were performed and pre-set values in GPS-X simulation software were applied. For the serial and parallel GP models used in hybrid model based on the default mechanistic model the same vector regressors were used, only the GP model hyperparameters were tuned to the default model predictions.

Fig. 5 shows the performance of the default mechanistic model prediction and the performance of a parallel-serial hybrid model, which was designed based on default mechanistic model. It can be seen that the prediction of the mechanistic model with the default model parameters (Fig. 5) is much worse compared to tuned mechanistic model (Fig. 3). This can be seen also in Fig. 6 where the distribution of the output-prediction error for the default model is presented. It can be seen that tuning of the model parameters results in the adjustment of the mean value of the error signal and slight variance correction.

The hybrid-model-output prediction based on default mechanistic model in Fig. 5 indicates that this model structure significantly improves the prediction. The output-prediction error and its distribution are similar to that obtained with the hybrid model based on the tuned mechanistic model.

Fig. 7 shows the output of the parallel GP model in the case of the tuned or default mechanistic model. It can be seen that the dynamics of both signals are similar. In the case of default mechanistic model, the parallel model structure also compensates for the off-set of the default mechanistic model.

Similar results were obtained also for effluent *TP* predictions (figures not shown).

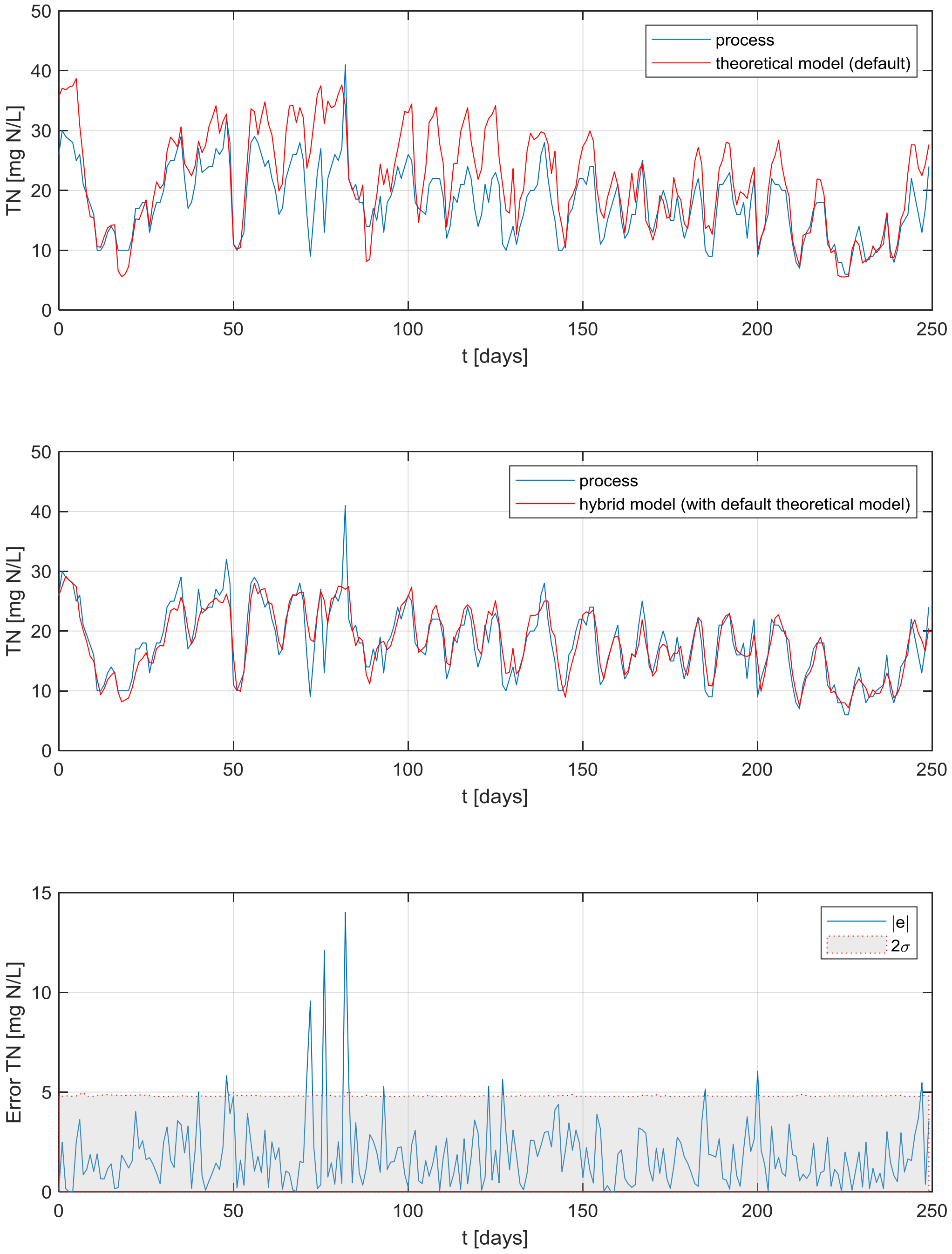

Fig. 5. Effluent total nitrogen (TN) prediction performed by the default mechanistic model (upper plot) and combined parallel–serial hybrid model (middle plot). The lower plot shows the absolute values of the error between the hybrid-model prediction and the true value, and 2 standard deviations of the data-driven model prediction (95 % confidence band).

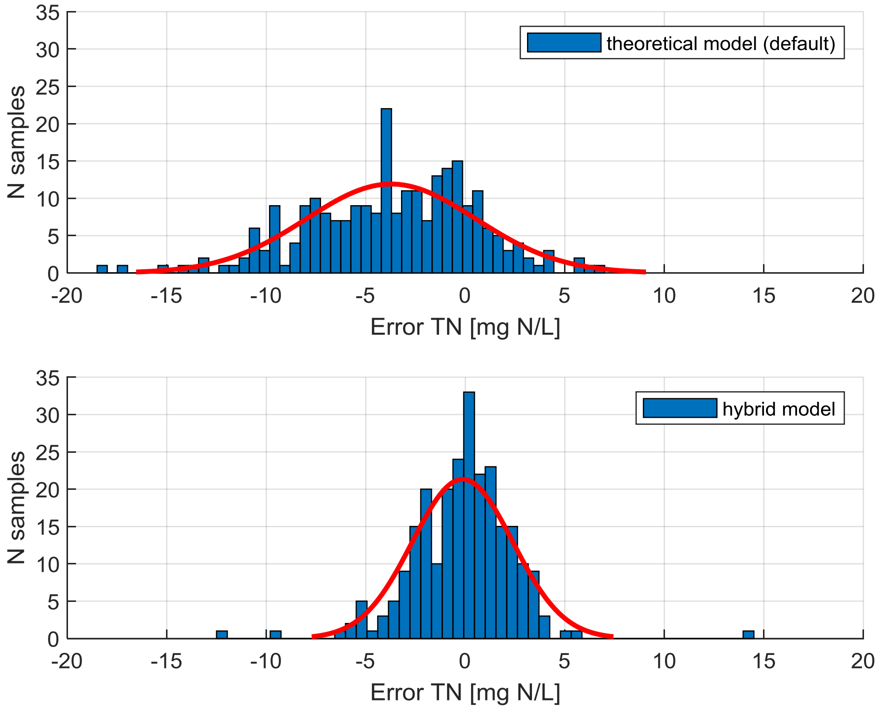

Fig. 6. Distribution of effluent total nitrogen (TN) model prediction error in the case of default mechanistic model and parallel-serial hybrid model structure based on default mechanistic model.

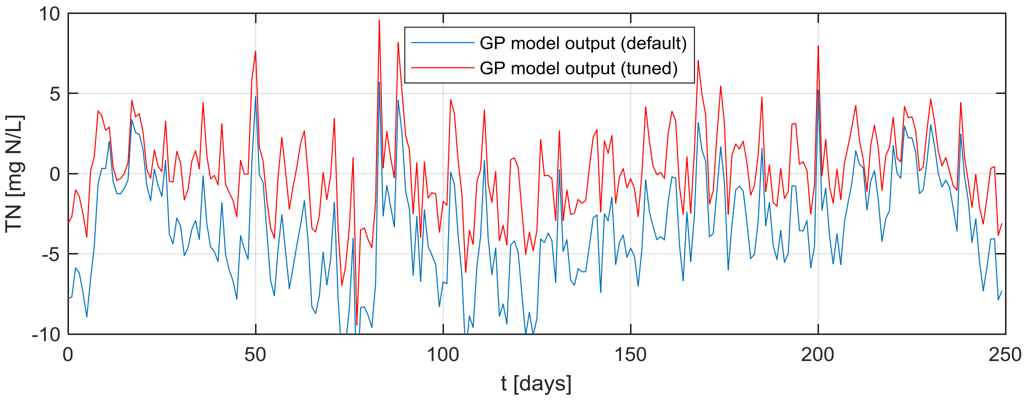

Fig. 7. The output of the parallel GP model in the hybrid model structure in the case of default and tuned mechanistic model.

4.4. Effluent *TN* and *TP* prediction accuracy

Comparison of evaluation criteria for different model structures in the case of effluent *TN* and *TP* predictions are presented in Table 4 and Table 5, respectively. The evaluation criteria confirm an improvement of effluent prediction by using a hybrid model structure. The improvement of both the serial hybrid model and the combined parallel-serial hybrid model compared to the mechanistic model is noticed. The improvement of the serial hybrid model is smaller and depends on case-specific data availability. In our case, the percentages of missing data for COD_i, TN_i, NH_4-N_i and TP_i were 8%, 20%, 10% and 21%, respectively.

From the results, it can be seen that the final output-prediction error is similar in the cases of a tuned or default mechanistic model used in the hybrid model structure. This indicates that difficult and time-consuming tuning of the mechanistic model parameters can be omitted in a hybrid model. The relative improvement of hybrid model structures compared to the mechanistic model is similar for *TN* or *TP*, but overall, the model quality for *TP* is worse compared to *TN*. This can be attributed to different reasons, e.g. lacking more detailed measurements for input *P* load as explained in Section 3.3, complex phosphorus kinetics that could not be captured by the hybrid model (Anderson et al., 2000), or unreliable prediction of phosphorus in ASM model if not accounting for corresponding physico-chemical processes (Kazadi Mbamba et al., 2016). It can be also noticed that with the highest values of R^2 equal to 0.81 and 0.67 for *TN* and *TP*, respectively, still a considerable proportion of the variance for the output variables could not be explained by the model. This may indicate that the available data is not sufficiently rich in the information or that important variables are not considered in the model (Côté et al., 1995). Further improvements of the model could be expected with additional monitoring or soft sensing data.

Table 4. Values of evaluation criteria for predicting effluent TN using different model options.

Effluent *TN*	**Default mechanistic model**		**Tuned mechanistic model**	
	R^2	**σ**	**R^2**	**σ**
Mechanistic model	0.065	4.270	0.619	3.620
Serial hybrid model	0.130	4.110	0.653	3.460
Parallel-serial hybrid model	0.814	2.526	0.808	2.563

Table 5. Values of evaluation criteria for predicting effluent TP and using different model options.

Effluent *TP*	**Default mechanistic model**		**Tuned mechanistic model**	
	R^2	**σ**	**R^2**	**σ**
Mechanistic model	0.110	0.766	0.542	0.789
Serial hybrid model	0.019	0.764	0.546	0.762
Parallel-serial hybrid model	0.668	0.695	0.651	0.710

4.5. The variance of the hybrid model prediction distribution

The GP model property of presenting the predicted model output with a distribution, expressed in terms of mean and variance, distinguishes this method from other data-driven methods. The mean value represents the most likely output and the variance can be interpreted as the measure of its confidence. The variance depends on the amount and quality of available data for prediction. To test how the variance changes with the change of data for prediction, some hypothetical examples were simulated. Simulation tests were performed on already presented test data by introducing offset changes in the parallel GP model input signals. The changes were introduced on TSS_i and $TN_{e,p}$ signals at t=150 d of the test data. The introduced offset change for $TN_{e,p}$ was small, i.e. within the identification data, while the change of TSS_i was significant, i.e. within two times of the identification data maximal value. It is expected that these artificially imposed signal changes on test data deteriorate the parallel model performance since they deviate from the training data.

Fig. 8 and Fig. 9 show the obtained results of the hybrid model predictions in both cases. In Fig. 8 it can be seen that after imposing the offset in $TN_{e,p}$, the hybrid-model prediction error increases and the model is no longer able to predict the process output close to the measured values. As can be seen in Fig. 8, the prediction variance and thus the 95% confidence band are not changed, which could be explained by the small $TN_{e,p}$ changes that are still within the range of identification data of the GP model. In the case of the offset change of TSS_i in Fig. 9, the introduced change is significant and all the TSS_i data are outside the identification region. It can be seen that in this case both the prediction error and the variance are increased, indicating deteriorated model performance and lower confidence in model prediction results.

The examples above indicate that the variance of the GP model predictions gives additional information on model prediction confidence. With the increased prediction variance due to the distance to identification data, it can be concluded that the model's confidence is decreased. It should be noted that in our full-scale case-study, the introduced change of GP model input signals to detect the increase of variance was quite large and needed to surpass the already high variance of the model due to available modelling data and corresponding measurement noise. A similar observation was noticed in (Samuelsson et al., 2017), where GP regression was used to detect a drift in an ammonium sensor (real data).

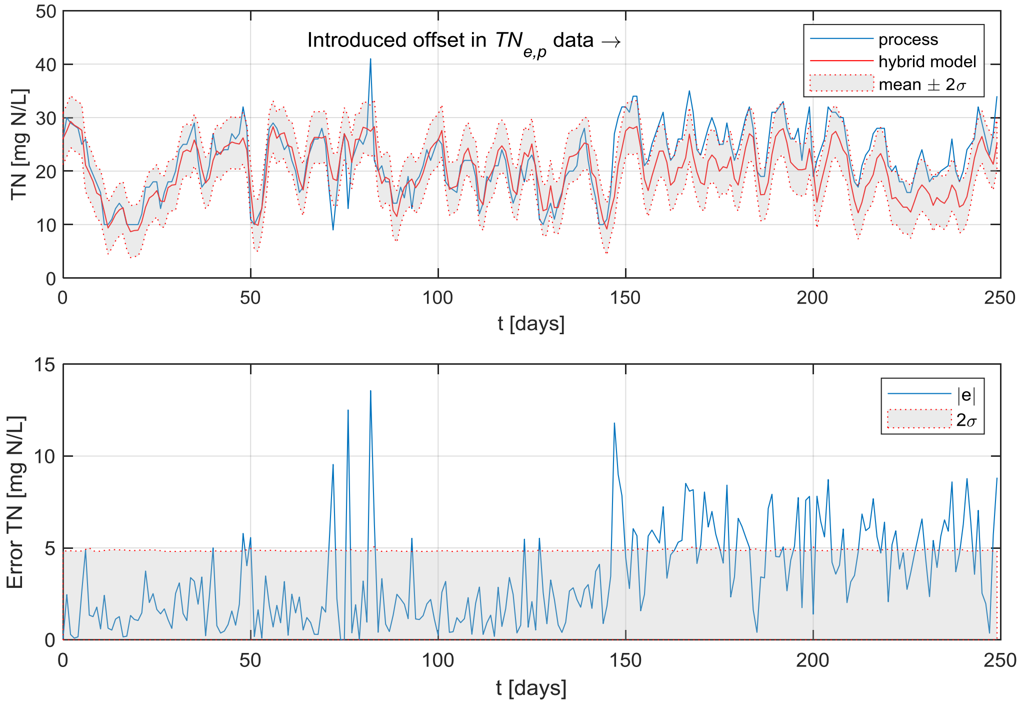

Fig. 8. Hybrid model performance on test data when an off-set is introduced on $TN_{e,p}$ signal at t=150 d. Deteriorated model prediction is noticed through the increased hybrid-model prediction error. The model variance and thus the 95% confidence band is not changed since the new data are still within the range of identification data.

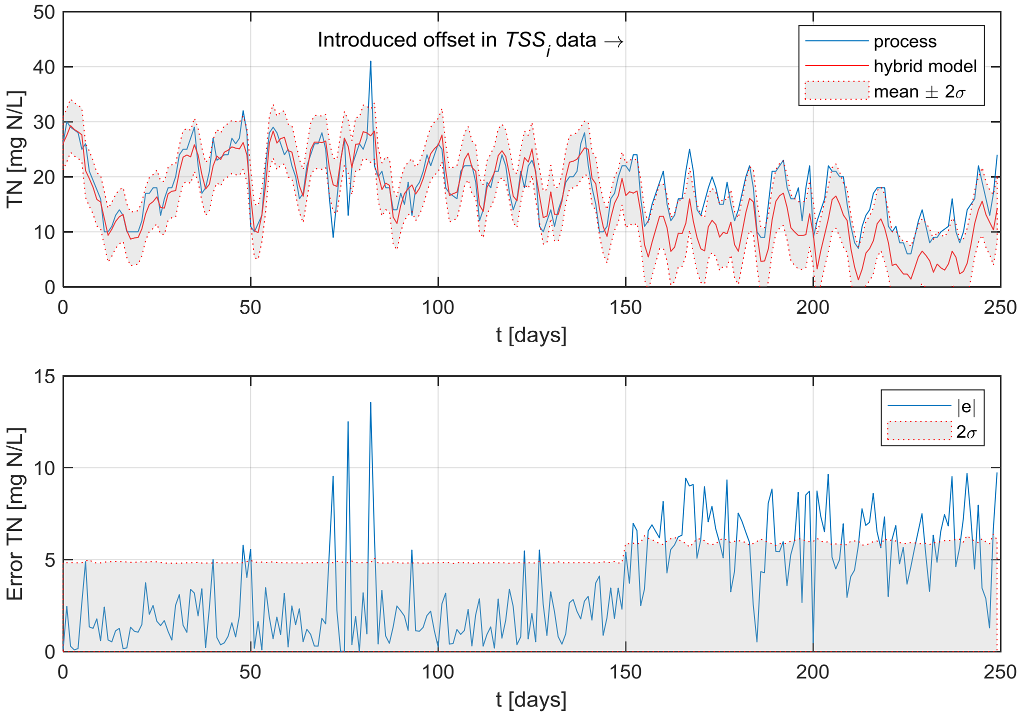

Fig. 9. Hybrid model performance on test data when an off-set is introduced on TSS_i signal at t=150 d. Deteriorated model prediction and lower confidence in model prediction results are noticed by the increased hybrid-model prediction error and variance. The variance and thus the 95% confidence band increased due to distance of new data to identification data.

5. Conclusions

This paper considers the design of hybrid mechanistic/GP models to be applied in WWTP modelling. Hybridisation is performed due to the lacking the mechanistic model, i.e. the widely applied mechanistic activated sludge model, to give accurate model predictions, as well as to address the gaps in input data that are crucial for the activated sludge model prediction.

The results show that both the serial and combined parallel-serial hybrid structures improve the model prediction accuracy in terms of statistical coefficient of determination R^2 and variance of the prediction error. The improvement of the serial structure, which fills the gaps of missing or erroneous input data, is smaller and depends on the case-specific data availability. The main contribution results from the parallel hybrid structure that models the mechanistic-model residuals. This indicates that despite the extensive effort in tuning mechanistic model parameters, there is still information latent in data that could not be tackled by the mechanistic model structure.

The most interesting results come from the application of the hybrid model based on the default mechanistic model and pre-set input wastewater characterization. The good prediction accuracy in this case, which is comparable to the prediction accuracy of hybrid model with tuned mechanistic-model parameters, indicates that no prior complex tuning of the mechanistic model is needed for good hybrid-model performance. Hence, the hybrid model could be used equally well without prior demanding wastewater characterization and adjustment of model parameters, which very much simplifies the hybrid-model design.

A realistic full-scale modelling study demonstrates the practical applicability of the approach. Further work will consider the application of the designed model in model-based control and optimisation.

Acknowledgement

This work was financially supported by the Slovenian Research Agency, program P2-0001 and research project L2-8174.

References

Anderson, J. S., McAvoy, T. J., Hao, O. J. (2000). Use of hybrid models in wastewater systems. Industrial & Engineering Chemistry Research, 39, 1694-1704.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer Science + Business Media.

Bradford, E., Schweidtmann, A. M., Zhang, D., Jing, K., Rio-Chanona, E. A. (2018). Dynamic modeling and optimization of sustainable algal production with uncertainty using multivariate Gaussian processes. Computers & Chemical Engineering, **118**, 143-158.

Chan, L. L. T., Chou, C., Chen, J. (2015). Hybrid model based control for membrane filtration process. IFAC Symposium on Dynamics and Control of Process Systems, including Biosystems, Trondheim, Norway.

Corominas, Ll., Garrido-Baserba, M., Villez, K., Olsson, G., Cortés, U., Poch, M. (2018). Transforming data into knowledge for improved wastewater treatment operation: A critical review of techniques. Environmental Modelling & Software, 106, 89-103.

Côté, M., Grandjean, B. P. A., Lessard, P., Thibault, J. (1995). Dynamic modelling of the activated sludge process: improving prediction using neural networks. Water Research, **29** (4), 995-1004.

Green, P. L., Cross, E. J., Worden, K. (2015). Bayesian system identification of dynamical systems using highly informative training data. Mechanical Systems and Signal Processing, **56–57**, 109-122.

Guo, H., Jeong, K., Lim, J., Jo, J., Kim, Y.M., Park, J., Kim, J.H., Cho, K.H. (2015). Prediction of effluent concentration in a wastewater treatment plant using machine learning models. Journal of Environmental Sciences, **32**, 90-101.

Haimi, H., Mulas, M., Corona, F., Vahala, R. (2013). Data-derived soft-sensors for biological wastewater treatment plants: an overview. Environmental Modelling & Software, **47**, 88-107.

Hajirahimi, Z., Khashei, M. (2019). Hybrid structures in time series modelling and forecasting: A review. Engineering Applications in Artificial Intelligence, **86**, 83-106.

Henze, M., Gujer, W., Mino, T., van Loosdrecht, M. (2000). Activated Sludge Models ASM1, ASM2, ASM2d and ASM3. IWA Publishing, London.

Hvala, N., Vrečko, D., Bordon, C. (2018). Plant-wide modelling for assessment and optimization of upgraded full-scale wastewater treatment plant performance. Water Practice & Technology, **13** (3), 566-582.

Hydromantis Environmental Software Solutions, Inc. (2016). GPS-X – Technical Reference, Version 6.5.

Kazadi Mbama, C., Flores-Alsina, X., Batstone, D. J., Tait, S. (2016). Validation of a plant-wide phosphorus modelling approach with minerals precipitation in a full-scale WWTP. Water Research, 100, 169-183.

Kocijan, J. (2016). Modelling and Control of Dynamic Systems Using Gaussian Process Models. Springer International Publishing, Cham.

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

Kruse, P. (2018). Review on water quality sensors. Journal of Physics D: Applied Physics, 51, 203002.

Lee, D. S., Jeon, C. O., Park, J. M., Chang, K. S. (2002). Hybrid neural network modeling of a full-scale industrial wastewater treatment process. Biotechnology and Bioengineering, **78** (6), 670-681.

Liu, Y., Xiao, H., Pan, Y., Huang, D., Wang, Q. (2016). Development of multi-step soft-sensors using Gaussian process model with application for fault prognosis. Chemometrics and intelligent laboratory systems, **157**(10), 85-95.

Mannina, G., Cosenza, A., Viviani, G. (2012). Uncertainty assessment of a model for biological nitrogen and phosphorus removal: Application to a large wastewater treatment plant. Physics and Chemistry of the Earth, 42-44, 61-69.

Manheim, D. C., Detwiler, R. L. (2019). Accurate and reliable estimation of kinetic parameters for environmental engineering applications: A global, multi objective, Bayesian optimization approach. MethodsX, **6**, 1398-1414.

Martin, C., Vanrolleghem, P. A. (2014). Analysing, completing, and generating influent data for WWTPmodelling: A critical review. Environmental Modelling & Software, 60, 188-201.

May, R., Dandy, G., Maier, H., 2011. Review of input variable selection methods for artificial neural networks, in: Suzuki, K. (Edt.), Artificial Neural Networks - Methodological Advances and Biomedical Applications. InTech, Rijeka, pp. 19–44.

Mulder, C. De, Flameling, T., Weijers, S., Amerlinck, Y., Nopens, I. (2018). An open software package for data reconciliation ad gap filling in preparation of water and resource recovery facility modeling. Environmental Modelling & Software, 107, 186-198.

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.

Peterka, V. (1981). Bayesian approach to system identification, in: Eykhoff, P. (Edt.), Trends and Progress in System Identification. Pergamon Press, Oxford, 239–304.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning. MIT press, Cambridge.

Samuelsson, O., Björk, A., Zambrano, J., Carlsson, B. (2017). Gaussian process regression for monitoring and fault detection of wastewater treatment processes. Water Science & Technology, **75** (12), 2952-2963.

Santín, I., Pedret, C., Vilanova, R. (2015). Effluent predictions in wastewater treatment plants for the control strategies selection. Actas de las XXXVI Jornadas de Automática, Bilbao, Spain.

Shi, J.Q., Choi, T., 2011. Gaussian process regression analysis for functional data. CRC Press, Boca Raton.

Takács, I., Patry, G. G., Nolasco, D. (1991). A dynamic model of the clarification-thickening process. Water Research, 25 (*10*) 1263-1271.

Vrečko, D., Hvala, N., Stražar, M. (2011). The application of model predictive control of ammonia nitrogen in an activated sludge process. Water Science & Technology, 64.5, 1115-1121.

Zambrano, J., Samuelsson, O., Carlsson, B. (2019). Machine learning techniques for monitoring the sludge profile in a secondary settler tank. Applied Water Science, 9, 146.