Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 54-7 (2021) 714–719

Integrated theoretical and data-driven Gaussian Process NARX Model for the Simulation of Effluent Concentrations in Wastewater Treatment Plant [★]

Tadej Krivec [*] **Nadja Hvala** [**] **Juš Kocijan** [***]

[*] *Jožef Stefan International Postgraduate School, Ljubljana, Slovenia*
and Jožef Stefan Institute, Ljubljana, Slovenia
(e-mail: tadej.krivec@ijs.si)
[**] *Jožef Stefan Institute, Ljubljana, Slovenia*
(e-mail: nadja.hvala@ijs.si)
[***] *Jožef Stefan Institute, Ljubljana, Slovenia*
and University of Nova Gorica, Nova Gorica, Slovenia
(e-mail: jus.kocijan@ijs.si)

Abstract: This paper presents the data-driven modelling part of a multi-input multi-output hybrid model of the Wastewater Treatment Plant (WWTP) and the simulation of the WWTP effluent. The information about the future effluent concentrations is important since it is used for efficient managing of the plant, for example in decision making, predictive control, quality control, and detection of violation of effluent limits. The hybrid model consists of a theoretical model based on first principles upgraded with a probabilistic data-driven model. The integrated model is based on a multi-input multi-output autoregressive Gaussian process (GP) model where the exogenous inputs include the predictions from the theoretical model. This approach allows us to use all available information in a single integrated model. We show significant improvement over the theoretical model for one-day-ahead prediction and validate the model for simulation, which can also be used when the effluent concentrations are not measured in an on-line fashion.

Keywords: Hybrid model, Gaussian process models, Nonparametric methods, Nonlinear system identification, Stochastic system identification, Simulation of stochastic systems

1. INTRODUCTION

Wastewater treatment is important for managing the quality of the water. Most often it is conducted in biological Wastewater Treatment Plants (WWTPs) aiming to convert and remove the pollutants to the level that treated water can be released to the water body or filtered. For efficient operation and control, a good model of the WWTP is required. Modelling and prediction of certain wastewater polluting compounds (e.g. organic matter, nitrogen, and phosphorus) is important in decision making because it helps to determine the acceptable impact of the WWTP effluent on the environment (Rahmat et al., 2011; Guo et al., 2015). Another aspect of using the model is to design energy-efficient control schemes of the plant (Vrečko et al., 2011).

A prediction of the effluent concentrations is usually obtained from a theoretical model that is derived from first principles. These theoretical models take into account the knowledge of the biological processes. Currently, the state-of-the-art theoretical models come from a family of activated sludge models (ASM) (Henze et al., 2000). Another approach is to obtain the predictions with a data-driven model which is based on directly measured process variables from sensors within the system and is usually modelled with multivariate statistics, fuzzy systems, or artificial neural networks (ANN).

The aforementioned approaches have their strengths and disadvantages. Theoretical models are time-consuming and require a lot of effort to adjust the model parameters to the real WWTP. Another limitation of the theoretical model is the lack of process knowledge that is yet to be explained. Data-driven models can suffer from the limited amount of process measurements. They are also bound to certain operating conditions and are less interpretable. On the other hand, the data-driven models can find patterns and deviations which can not (yet) be explained by a theoretical model. The hybrid model takes into account both the benefits of the aforementioned approaches, i.e., the prior knowledge of the physical process and data-driven insight from measured data. The prediction from the theoretical model is used as an input to the data-driven model. This results in an integrated model that can model the remaining deviations from the current theoretical model based on the sensor measurements. A hybrid approach where the ANNs are used for the data-driven part can be found in (Anderson et al., 2000; Lee et al.,

[★] The authors acknowledge the research core funding No. P2-0001, and Ph.D. grant for Tadej Krivec which were financially supported by the Slovenian Research Agency.

10.1016/j.ifacol.2021.08.445

2002; Cote et al., 1995), and a probabilistic approach based on Gaussian processes (GPs) can be found in (Hvala and Kocijan, 2020).

ANNs are prone to overfit given the small dataset available. Therefore, it is difficult to identify a good model with ANNs, mainly because of the many metaparameters involved. Another drawback is that the uncertainty of the system is not accounted for. Unlike ANNs, GPs proved to work well on smaller datasets, can model the uncertainty in the system, and protect well against overfitting since the Bayesian nonparametric approach penalizes overly complex models. The information about the predicted uncertainty can also help to assess the violations of the conditions under which the theoretical model has been retrieved. The limitation of the current GP models for modelling WWTP effluent is that they use multiple multi-input single-output models, which can not provide multi-day-ahead predictions within a single integrated model (in an autoregressive manner).

This paper presents a hybrid model based on a multi-input multi-output Gaussian Process Nonlinear AutoRegressive eXogenous (GP-NARX) model which takes into account all the available information in a single integrated model. Our contributions are the following:

- We consider a hybrid multi-input multi-output GP-NARX model to simultaneously model the output variables of interest inside a single integrated model for the effluent concentrations in WWTP.
- We extend the software framework for modelling with GPs (van der Wilk et al., 2020) for multi-output training and simulation of GP-NARX models.
- We improve on the previous work on hybrid GP-NARX models for WWTP, where we not only consider prediction, but also the simulation of the effluent concentrations.

The remaining of the paper is structured as follows. In Section 2 we present a multi-input multi-output GP-NARX model and propose a simple (but efficient) sampling scheme for the effluent simulation. In Section 3 we consider the WWTP case study and present the results of the prediction and the simulation of the effluent concentrations. Lastly, we discuss the limitations of the current work, future work, and conclude with the final remarks in Section 4.

2. MULTI-INPUT MULTI-OUTPUT GP-NARX MODEL

The process is described by

$$\mathbf{Y}_{i,p} = f^p(\mathbf{Z}_{i,:}) + \epsilon_i^p, \tag{1}$$

where $\mathbf{Y} \in \mathcal{R}^{n \times p}$ represents a multi-dimensional matrix of the observed outputs and the mapping f represents a nonlinear mapping modeled with a GP. The latent function f^p is presumed to be corrupted with independent and identically distributed noise that follows a Gaussian distribution $\epsilon_i^p \sim \mathcal{N}(\mathbf{0}, \sigma_p^2)$. The input matrix $\mathbf{Z} \in \mathcal{R}^{n \times (p \cdot n_a + d \cdot n_b)}$ is represented with a Nonlinear AutoRegressive eXogenous (NARX) model. The i-th row of the input matrix $\mathbf{Z}$ is defined by

$$\mathbf{Z}_{i,:} = [\mathbf{Y}_{i-1,:}, \dots, \mathbf{Y}_{i-n_a,:}, \mathbf{X}_{i,:}, \dots, \mathbf{X}_{i-n_b,:}], \tag{2}$$

where $\mathbf{Y}_{i,:}$ defines the i-th row of the output matrix $\mathbf{Y}$ and $[\mathbf{X}_{i,:}, \dots, \mathbf{X}_{i-n_b,:}] \in \mathcal{R}^{1 \times d \cdot n_b}$ the i-th row of the matrix with exogenous inputs. Meta-parameters n_a and n_b denote the number of lags.

Let the matrix $\mathbf{F}$ represent the matrix of latent function values where $\mathbf{F}_{i,p} = f^p(\mathbf{Z}_{i,:})$. Let $\mathbf{f}^p = \mathbf{F}_{:,p}$ and $\mathbf{f} = [\mathbf{f}^{1^T}, \dots, \mathbf{f}^{p^T}]^T \in \mathcal{R}^{n \cdot p \times 1}$ and let the same notation hold for $\mathbf{Y}$. A GP then defines the prior over the vector of latent function values $\mathbf{f}$ (Rasmussen and Williams, 2006). Allowing separate parametrizations of each f^p, a GP is fully specified by a separate mean $m^p(\mathbf{Z}_{i,:})$ and a separate covariance function $k^p(\mathbf{Z}_{i,:}, \mathbf{Z}_{j,:})$ for each output p. The covariance function has to satisfy the condition of generating a semi-positive definite matrix. Many popular choices can be found in (Kocijan, 2016), where combinations of the covariance functions are also permitted. Without loss of generality, we select the mean function as 0.

Let θ^p define the hyperparameters of the covariance function k^p. The covariance matrix of the vector of latent function values $p(\mathbf{f}^p|\mathbf{Z}, \theta^p) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{ff}^p)$ is defined by

$$\mathbf{K}_{ff}^p = \begin{bmatrix} k^p(\mathbf{Z}_{1,:}, \mathbf{Z}_{1,:}) & \dots & k^p(\mathbf{Z}_{1,:}, \mathbf{Z}_{n,:}) \\ \vdots & \ddots & \vdots \\ k^p(\mathbf{Z}_{n,:}, \mathbf{Z}_{1,:}) & \dots & k^p(\mathbf{Z}_{n,:}, \mathbf{Z}_{n,:}) \end{bmatrix}. \tag{3}$$

The likelihood of $\mathbf{f}$ fully specifies the probabilistic model. Allowing separate noise levels for each column it is defined by $p(\mathbf{y}^p|\mathbf{f}^p, \sigma_p^2) = \mathcal{N}(\mathbf{f}^p, \mathbf{K}_{ff}^p + \sigma_p^2\mathbf{I})$. The covariance of the joint prior distribution over all vectors of the observed values $p(\mathbf{y}|\mathbf{Z}, \theta, \sigma^2) \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{yy})$ is then defined by

$$\mathbf{K}_{yy} = \begin{bmatrix} \mathbf{K}_{ff}^1 + \sigma_1^2\mathbf{I} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{K}_{ff}^2 + \sigma_2^2\mathbf{I} & \dots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{K}_{ff}^p + \sigma_p^2\mathbf{I} \end{bmatrix}, \tag{4}$$

where $\theta = \{\theta^1, \dots, \theta^p\}$ and $\sigma^2 = \{\sigma_1^2, \dots, \sigma_p^2\}$. The block-diagonal structure implies that the outputs are conditionally independent given the inputs and hyperparameters. Assuming $\mathbf{f}_{t+1}$ represents a vector of latent function values at the next time step t and the corresponding input is denoted by $\mathbf{z}_{t+1}$, the posterior distribution over the vector of latent function values is defined by

$$p(\mathbf{f}, \mathbf{f}_{t+1}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+1}, \theta, \sigma^2) = \frac{p(\mathbf{y}|\mathbf{f}, \sigma^2)p(\mathbf{f}, \mathbf{f}_{t+1}|\mathbf{Z}, \mathbf{z}_{t+1}, \theta)}{p(\mathbf{y}|\mathbf{Z}, \theta, \sigma^2)}. \tag{5}$$

2.1 Hyperparameter estimation

Hyperparameters θ and σ^2 can be determined with the maximization of the marginal-log-likelihood defined by

$$\log p(\mathbf{y}|\mathbf{Z}, \theta, \sigma^2) = -\frac{1}{2}\log(|\mathbf{K}_{yy}|) - \frac{1}{2}\mathbf{y}^T\mathbf{K}_{yy}^{-1}\mathbf{y} - \frac{np}{2}\log(2\pi). \tag{6}$$

Since the covariance matrix $\mathbf{K}_{yy}$ is block-diagonal, each block can be parametrized with a separate set of hyperparameters and different covariance function choices. The

We use $\mathbf{A}_{i,:} \in \mathcal{R}^{1 \times n}$ to represent the i-th row of the matrix $\mathbf{A}$, whereas $\mathbf{A}_{:,j} \in \mathcal{R}^{m \times 1}$ represents the j-th column of the matrix $\mathbf{A} \in \mathcal{R}^{m \times n}$.

number of hyperparameters can be reduced by estimating a single covariance function with only one set of hyperparameters. Gradients of the objective can be analytically obtained (Rasmussen and Williams, 2006). Hereafter we omit the conditional dependency on θ and σ^2 in the notation for convenience.

2.2 Prediction

Predictive distribution is obtained by integrating the latent posterior out of the likelihood at test inputs

$$p(\mathbf{y}_{t+1}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+1}) = \int p(\mathbf{y}_{t+1}|\mathbf{f}_{t+1}) p(\mathbf{f}_{t+1}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+1}) d\mathbf{f}_{t+1}. \quad (7)$$

The mean and the variance of the predictive distribution can be evaluated in closed-form and are defined by

$$\mu_{t+1}(p(\mathbf{y}_{t+1}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+1})) = \mathbf{K}_{y_* y} \mathbf{K}_{yy}^{-1} \mathbf{y}, \quad (8a)$$

$$\sigma^2_{t+1}(p(\mathbf{y}_{t+1}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+1})) = \mathbf{K}_{y_* y_*} - \mathbf{K}_{y_* y} \mathbf{K}_{yy}^{-1} \mathbf{K}_{yy_*} \quad (8b)$$

where $\mathbf{K}_{y_* y_*}$, $\mathbf{K}_{y_* y}$, and $\mathbf{K}_{yy}$ denote the covariance matrices between the test inputs $\mathbf{z}_{t+1}$, between the test and training inputs $\mathbf{Z}$, and between training inputs $\mathbf{Z}$ respectively.

2.3 Simulation

Simulation is obtained in the form of a nonlinear output error (NOE) model. The first step (i.e., $t+1$) is identical to the prediction defined with the equation (7). At the time step $t+2$ the input vector to the distribution $p(\mathbf{y}_{t+2}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+2})$ is defined by

$$\mathbf{z}_{t+2} = [\mathbf{f}^T_{t+1}, \dots, \mathbf{f}^T_{t-n_a}, \mathbf{X}_{t+2,:}, \dots, \mathbf{X}_{t+2-n_b,:}], \quad (9)$$

where $\mathbf{f}_{t+1}$ follows a multivariate Gaussian distribution (i.e., the latent predictive distribution $p(\mathbf{f}_{t+1}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+1})$). The input $\mathbf{z}_{t+2}$ is therefore no longer deterministic, but rather an uncertain input and makes the integral

$$\int p(\mathbf{y}_{t+2}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+2}) p(\mathbf{z}_{t+2}) d\mathbf{z}_{t+2} \quad (10)$$

intractable. The predictive distribution at the time step $t+2$ is approximated in the form of a Gaussian Mixture Model (GMM) defined by

$$p(\mathbf{y}_{t+2}|\mathbf{y}, \mathbf{Z}, \mathbf{z}_{t+2}) \approx \frac{1}{m} \sum_{i=1}^{m} p(\mathbf{y}_{t+2}|\mathbf{y}, \mathbf{Z}, \hat{\mathbf{z}}^i_{t+2}), \quad (11)$$

where

$$\hat{\mathbf{z}}^i_{t+2} = [\hat{\mathbf{f}}^{iT}_{t+1}, \dots, \mathbf{f}^T_{t-n_a}, \mathbf{X}_{t+2,:}, \dots, \mathbf{X}_{t+2-n_b,:}], \quad (12)$$

and $\hat{\mathbf{f}}^i_{t+1}$ is a sample from the latent predictive distribution. The number of samples is denoted with m. This process can be repeated up to an arbitrary time step into the future where the samples of $\hat{\mathbf{f}}^i_{t+q}$ are drawn from a GMM instead of a Gaussian distribution for all $q \geq 2$.

3. CASE STUDY

The WWTP plant considered in this study is subjected to control which heavily depends on the accurate model of the plant. This case study focuses on the simulation model of nitrogen and phosphorus in an autoregressive manner which enables the use of historic measurements for estimating the model parameters, and iterative prediction even when the nitrogen and phosphorus concentrations are not measured on-line.

3.1 WWTP description

The plant consists 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 a sludge treatment (sludge thickening, aerobic digestion, dewatering, and sludge drying). To design efficient and reliable control schemes, a good predictive model for nitrogen and phosphorus is needed. The data used in this study are presented in Table 1 where Q_i stands for influent flow, COD_i for influent chemical oxygen demand, TN_i for influent total nitrogen, $\mathrm{NH4}_i$ for influent ammonia nitrogen, TP_i for influent total phosphorus, T_w for wastewater temperature, DO_{b1} for dissolved oxygen concentration at the beginning of the aeration tank, DO_{b2} for dissolved oxygen concentration at the end of the aeration tank, Q_r for return sludge flow, MLSS_b for mixed liquor suspended solids concentration at the outlet of the aeration tank, Q_{rw} for reject water flow, TN_{th} for effluent total nitrogen predicted by the theoretical model, TP_{th} for effluent total phosphorus predicted by the theoretical model, TN_p for measured effluent total nitrogen, and TP_p for measured effluent total phosphorus. The data consists of 650 daily measurements, where the first 400 measurements are selected for the training dataset and the rest as a test dataset.

3.2 WWTP theoretical model

For predicting the nitrogen and phosphorus effluent concentrations, a theoretical model was first proposed. Table 1 shows the measured outputs $\mathbf{Y}^{th}$ and the measured inputs $\mathbf{X}^{th}$ that were used to tune the theoretical model. The measured outputs are only used for tuning the model and are not used as historic inputs for predicting the concentrations of the effluent. The theoretical model can therefore

Table 1. Data used in the case study for modelling the nitrogen and phosphorus concentrations in a full-scale WWTP plant. Matrix $\mathbf{X}^{th}$ represents the inputs to the theoretical model and $\mathbf{Y}^{th}$ the measurements of the outputs to which the theoretical model is tuned. Matrix $\mathbf{X}$ is the matrix of inputs which together with the delayed output measurements $\mathbf{Y}$ parametrizes the inputs to the GP-NARX model.

	Theoretical model		GP-NARX model	
	$\mathbf{X}^{th}$	$\mathbf{Y}^{th}$	$\mathbf{X}$	$\mathbf{Y}$
Q_i	×	-	×	-
COD_i	×	-	×	-
TN_i	×	-	×	-
$\mathrm{NH4}_i$	×	-	×	-
TP_i	×	-	×	-
T_w	×	-	×	-
DO_{b1}	×	-	×	-
DO_{b2}	×	-	×	-
Q_r	×	-	×	-
MLSS_b	×	-	×	-
Q_{rw}	-	-	×	-
TN_{th}	-	-	×	-
TP_{th}	-	-	×	-
TN_p	-	×	-	×
TP_p	-	×	-	×

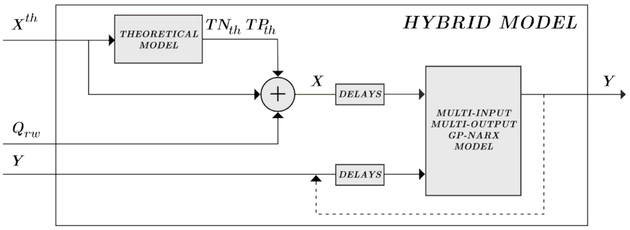

Fig. 1. The hybrid model used for WWTP plant modelling. The exogenous inputs to the Gaussian process model are the same inputs as to the theoretical model $\mathbf{X}^{th}$ augmented with the predictions from the theoretical model TN_{th} and TP_{th}, additional measurement Q_{rw} and all their laggs as shown in Table 1. The solid lines represent the model for parameter estimation. The dashed line represents the back-propagated outputs in simulation. In that case, the delayed outputs are random variables, rather than deterministic measurements.

also be seen as a simulation model since the predicted effluent concentrations at an arbitrary time step in the future only depend on the measurements of the input for the time step considered. The theoretical model of the plant was designed and tuned with plant measurements. The aerobic reactors in the biological stage are modelled with ASM2d model. The secondary clarifiers are considered biologically inactive and are modelled with the double exponential settling velocity function (Takács et al., 1991). The full model is developed in GPS-X simulation software (Hydromantis, 2016). The details of the theoretical model can be found in (Hvala et al., 2018).

3.3 Hybrid model

For the data-driven part of a hybrid model, a multi-input multi-output Gaussian process was used, which was described in Section 2. The hybrid model considered in this case study is shown on Figure 1. The data used for estimating the model hyperparameters are presented in Table 1. We can see that the theoretical knowledge of the process is introduced to the data-driven model with the inclusion of the theoretical predictions of nitrogen (TN_{th}) and phosphorus (TP_{th}) in the input matrix $\mathbf{X}$ which parametrizes the inputs to a GP-NARX model.

Table 2. The results of a 10-fold cross-validation on the training dataset for hyperparameter estimation of shared (S.) and independent (I.) covariance functions (k), where TN_{hyb} and TP_{hyb} represent the one-day-ahead predictions from the hybrid model of nitrogen and phosphorus concentrations respectively. The final choice of covariance functions is shown in bold.

	TN_{hyb}		TP_{hyb}	
	k^1	SMSE	k^2	SMSE
I.	Lin. + M52$_A$	0.244 ± 0.12	Lin. + M52$_A$	0.317 ± 0.11
I.	Lin. + RBF$_A$	0.249 ± 0.12	Lin. + RBF$_A$	0.316 ± 0.10
I.	**Lin. + M52**	**0.210 ± 0.11**	Lin. + M52	0.297 ± 0.09
I.	Lin. + RBF	0.211 ± 0.11	**Lin. + RBF**	**0.296 ± 0.10**
S.	Lin. + M52$_A$	0.214 ± 0.09	Lin. + M52$_A$	0.320 ± 0.10
S.	M52$_A$	0.244 ± 0.15	M52$_A$	0.347 ± 0.20
S.	Lin. + RBF$_A$	0.218 ± 0.11	Lin. + RBF$_A$	0.329 ± 0.11
S.	RBF$_A$	0.238 ± 0.14	RBF$_A$	0.350 ± 0.21
S.	Lin. + M52	0.204 ± 0.09	Lin. + M52	0.310 ± 0.09
S.	M52	0.223 ± 0.11	M52	0.315 ± 0.11
S.	Lin. + RBF	0.202 ± 0.09	Lin. + RBF	0.312 ± 0.10
S.	RBF	0.217 ± 0.11	RBF	0.314 ± 0.11

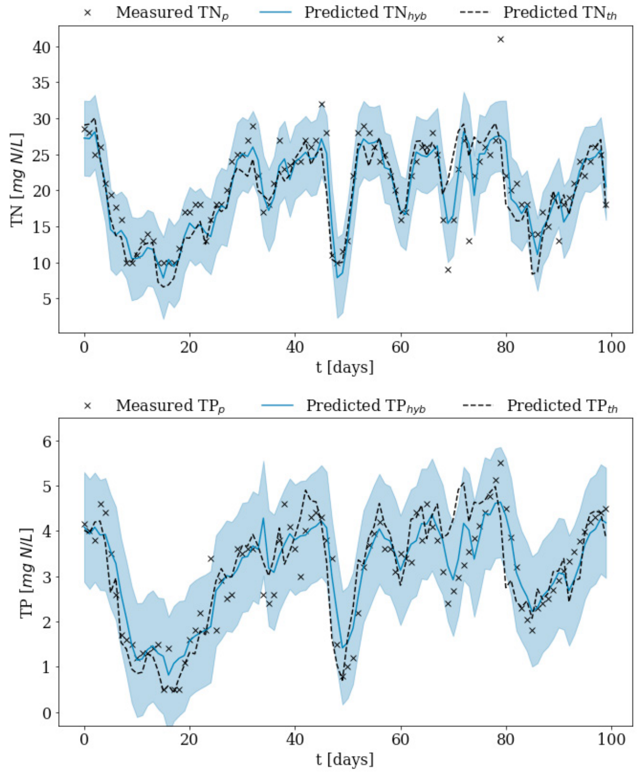

Fig. 2. Comparison of a hybrid one-day-ahead prediction (TN_{hyb}, TP_{hyb}), and a theoretical one-day-ahead prediction (TN_{th}, TP_{th}), for the first 100 days on the test dataset. The solid region represents a 2 standard deviation interval of the prediction.

Block diagonal joint distribution of a GP-NARX model presented with equation (4) allows us to estimate the hyperparameters of the model with 2 schemes, either a covariance function is shared among the block elements and only one set of hyperparameters is optimized, or each block is parameterized by a separate covariance function.

Multiple covariance functions and their combinations from (Kocijan, 2016) were tested empirically through 10-fold cross-validation on the training dataset, where Table 2 only shows some on them. In the Table 2 Lin. stands for linear covariance function, RBF for radial basis function, M52 for Matérn52 covariance function, and the subscript in, e.g. $M52_A$, denotes the Automatic Relevance Determination (ARD) property. Covariance functions used in Table 2 are defined in the Appendix A. We can see in Table 2 that the best overall results were obtained from hyperparameter estimation of independent covariance functions for each output, where a combination of a linear and Matérn52 covariance function was used for modelling the nitrogen concentration and a combination of a linear and RBF covariance function for modelling the phosphorus concentration.

The objective, defined by equation (6), was optimized with Adam (Kingma and Ba, 2014) with learning rate α =

Table 3. Comparison of the performance between the hybrid-model prediction (TN_{hyb}, TP_{hyb}), and theoretical-model prediction (TN_{th}, TP_{th}), on the test dataset. The best results are shown in bold.

	TN_{hyb}	TP_{hyb}	TN_{th}	TP_{th}
SMSE	**0.180**	**0.229**	0.254	0.378
PCC	**0.902**	**0.879**	0.869	0.811
MSLL	-2.50	-1.00	-	-

Table 4. Comparison of the performance between the hybrid-model simulation (TN_{hyb}, TP_{hyb}), and theoretical-model simulation (TN_{th}, TP_{th}), on the test dataset. The best results are shown in bold.

	TN_{hyb}	TP_{hyb}	TN_{th}	TP_{th}
SMSE	0.261	**0.281**	**0.254**	0.378
PCC	0.856	**0.849**	**0.869**	0.811
MSLL	-2.067	-0.907	-	-

0.1 and 500 iterations. Optimization parameters and the autoregressive parameters from equation (2) were found empirically over 10-fold cross-validation. Autoregressive parameters were selected as $[n_a,\ n_b] = [1,\ 2]$. Increasing the lags did not significantly improve the results.

3.4 Results

In this section, we present the results for the one-day-ahead prediction and the simulation of nitrogen and phosphorus effluent concentrations in the WWTP.

Prediction Figure 2 shows the first 100 days of the predicted response for nitrogen and phosphorus concentrations on a test dataset. The whole training dataset was used for estimating the hyperparameters, where the metaparameters were choosen as previously described in Section 3.3. We can see that the concentrations are well described with the predicted means. Also, the shaded region of 2 standard deviations captures the measured values well. Table 3 shows the comparison between the hybrid-model prediction and the theoretical-model prediction for 2 standard performance measures, standardized mean squared error (SMSE) (Kocijan, 2016) and Pearson correlation coefficient (PCC) (Freedman et al., 2007). Mean standardized log loss (MSLL) is also presented, which is more suitable for validation in the form of random variables, weighting the error by the predicted standard deviation. MSLL is approximately zero for simple methods and negative for better methods and is defined in Appendix B. We can see in Table 3 that the effluent predictions from the theoretical model are significantly improved with the hybrid model. Also, MSLL shows that not only the means are well modeled, but also the predicted uncertainty.

Simulation The ultimate model validation is the simulation. This can be seen as training a NARX model, and validating the learned hyperparameters with a NOE model. Figure 3 shows 50 independent free-run simulation responses of nitrogen and phosphorus concentrations on the test dataset and the corresponding means of the samples. We can see that the simulation samples describe the measured variables well. Table 4 shows the SMSE, PCC, and MSLL for the simulation response, where the distribution at each step is approximated with a Gaussian from 300 independent simulation samples (50 samples are only used for the purpose of presentation). Similarly, as with the prediction, MSLL shows that the concentrations are well modelled in terms of probability distributions. Overall, we can see that the hybrid model achieves better results than the theoretical one. Worse SMSE for the simulation of the concentration of nitrogen is not significant in comparison with a far better SMSE for the simulation of the concentration of phosphorus.

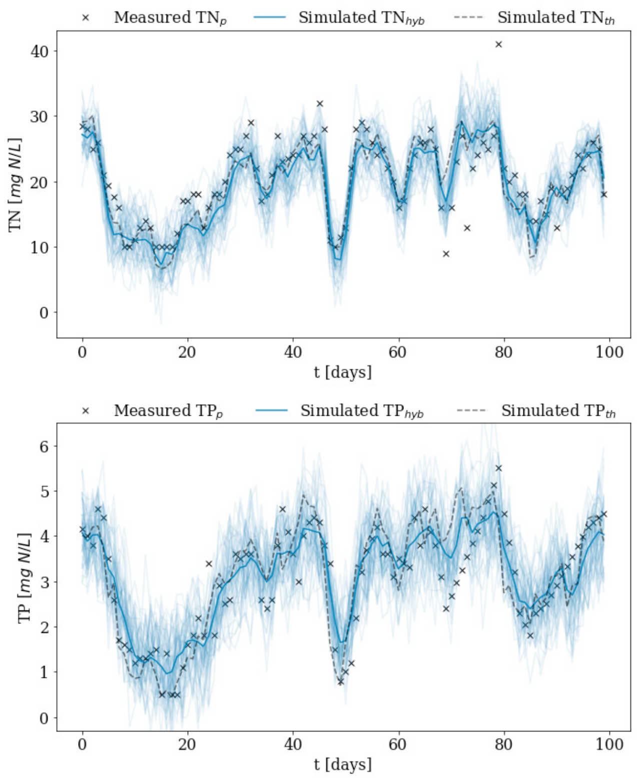

Fig. 3. Comparison of a hybrid simulation (TN_{hyb}, TP_{hyb}), and theoretical simulation (TN_{th}, TP_{th}), for the first 100 days on the test dataset. The thick blue line represents the mean of the 300 simulated samples.

4. CONCLUSION

This paper presented the data-driven modelling part of a hybrid model of WWTP. The managing of wastewater is essential in water quality control and has a significant environmental impact. Accurate predictions of nitrogen and phosphorus concentrations are important in devising efficient control schemes, which aim to satisfy the acceptable impact of the WWTP effluent on the environment. To devise optimal control schemes, both multiple-day-ahead prediction and its uncertainty estimation are heavily desired.

Our modelling solution significantly improved the predictions of nitrogen and phosphorus concentrations compared to the existing theoretical approach. The existing hybrid approach using GPs consisted of multiple multi-input single-output models, whereas we considered a single multi-input multi-output GP model. This allows us to concurrently simulate multiple outputs considered in this study up to the desired horizon. From the results, we conclude that our model can be also used for simulation, which is essential in the case when the nitrogen and phosphorus concentrations are not measured on-line.

The current limitation of our approach is that the joint distribution of GP-NARX model is block-diagonal, which only considers conditionally independent outputs given the inputs and hyperparameters. A future improvement would be to consider a non-block-diagonal joint distribution (Guo et al., 2010). Another limitation of our approach is that the hyperparameters of the GP-NARX model are estimated for prediction, where we do not consider simulation in their estimation. This could be improved with models that do consider simulation, e.g., NOE models or state-space models.

REFERENCES

Anderson, J., McAvoy, T., and Hao, O. (2000). Use of hybrid models in wastewater systems. *Industrial & Engineering Chemistry Research*, 39(6), 1694–1704.

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

Freedman, D., Pisani, R., and Purves, R. (2007). Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York.*

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

Guo, S., Sanner, S., and Bonilla, E.V. (2010). Gaussian process preference elicitation. In *Advances in neural information processing systems*, 262–270.

Henze, M., Gujer, W., Mino, T., and van Loosdrecht, M.C. (2000). *Activated sludge models ASM1, ASM2, ASM2d and ASM3.* IWA publishing.

Hvala, N. and Kocijan, J. (2020). Design of a hybrid mechanistic/gaussian process model to predict full-scale wastewater treatment plant effluent. *Computers & Chemical Engineering*, 106934.

Hvala, N., Vrečko, D., and 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 (2016). Hydromantis environmental software solutions, inc.

Kingma, D.P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

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

Lee, D.S., Jeon, C.O., Park, J.M., and Chang, K.S. (2002). Hybrid neural network modeling of a full-scale industrial wastewater treatment process. *Biotechnology and bioengineering*, 78(6), 670–682.

Rahmat, M., Samsudin, S., Wahab, A.P.I.D.N., Sy Salim, S.N., and Gaya, M. (2011). Control strategies of wastewater treatment plants. *Australian Journal of Basic and Applied Sciences*, 5, 446–455.

Rasmussen, C.E. and Williams, C.K. (2006). *Gaussian processes for machine learning.* MIT press Cambridge.

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

van der Wilk, M., Dutordoir, V., John, S., Artemev, A., Adam, V., and Hensman, J. (2020). A framework for interdomain and multioutput gaussian processes.

Vrečko, D., Hvala, N., and Stražar, M. (2011). The application of model predictive control of ammonia nitrogen in an activated sludge process. *Water Science and Technology*, 64(5), 1115–1121.

Appendix A. COVARIANCE FUNCTIONS

A.1 Linear covariance function

Linear covariance function is defined by

$$k(\mathbf{Z}_{i,:}, \mathbf{Z}_{j,:}) = \sigma_f^2 \mathbf{Z}_{i,:} \mathbf{Z}_{j,:}, \tag{A.1}$$

where σ_f denotes a scaling factor.

A.2 Radial basis covariance function

Radial basis covariance function is defined by

$$k(\mathbf{Z}_{i,:}, \mathbf{Z}_{j,:}) = \sigma_f^2 e^{-\frac{1}{2}r^2}, \tag{A.2}$$

where $r = ||\frac{1}{l}(\mathbf{Z}_{i,:} - \mathbf{Z}_{j,:})||$ and l represents a lengthscale parameter. Automatic Relevance Determination (ARD) property weights the columns of the input $\mathbf{Z}$ with their corresponding lengthscale l_d, where r is defined by

$$r = \sqrt{(\mathbf{Z}_{i,:} - \mathbf{Z}_{j,:})^T \mathbf{\Lambda}^{-1} (\mathbf{Z}_{i,:} - \mathbf{Z}_{j,:})}, \tag{A.3}$$

and $\mathbf{\Lambda}^{-1} = \text{diag}([l_1^{-2}, \ldots, l_d^{-2}])$, where d is the number of columns in matrix $\mathbf{Z}$.

A.3 Matérn52 covariance function

A Matérn52 covariance function is defined by

$$k(\mathbf{Z}_{i,:}, \mathbf{Z}_{j,:}) = \sigma_f^2 (1 + \sqrt{5}r + \frac{5}{3}r^2) e^{-\sqrt{5}r}. \tag{A.4}$$

Matérn52 covariance function with an ARD property defines r the same as in the Radial basis covariance function with equation A.3.

Appendix B. PERFORMANCE MEASURES

B.1 Mean Standardized Log Loss

The mean standardized log loss (MSLL) is defined by

$$\begin{aligned} \text{MSLL} = \frac{1}{2N} \sum_{i=1}^{N} \left[\ln(\sigma_i^2) + \frac{(y_i - \mu_i))^2}{\sigma_i^2} \right] \\ - \frac{1}{2N} \sum_{i=1}^{N} \left[\ln(\mathbb{V}(\mathbf{y})) + \frac{(y_i - \mathbb{E}(\mathbf{y}))^2}{\mathbb{V}(\mathbf{y})} \right], \end{aligned} \tag{B.1}$$

where $\mathbf{y}$ represents the ground truth, μ_i the predicted mean at time step i, σ_i^2 the predictive variance at time step i, and N the number of data points.