Young Water Professionals Regional Conference
the international water association

Combining mechanistic and data-driven techniques for predictive modelling of wastewater treatment plants

W., Quaghebeur[1,2,*], B., De Jaeger[1], B., De Baets[2], I., Nopens[1]

[1] BIOMATH, Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, B-9000 Gent, Belgium
[2] KERMIT, Department of Data Analysis and Mathematical Modelling, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, B-9000 Gent, Belgium
[*] ward.quaghebeur@ugent.be

Abstract: Mechanistic models are widely used for modelling of wastewater treatment plants. However, as they are based on simplified and incomplete domain knowledge, they often lack accurate predictive capabilities. In contrast, data-driven models are able to make accurate predictions, but only in the operational regions that are sufficiently described by the dataset used. We investigate an alternative hybrid model, combining mechanistic and data-driven techniques. We show that the hybrid approach combines the strengths of both modelling paradigms. It allows for accurate predictions out of the training dataset without the need for complete domain knowledge. Moreover, this approach is not limited to wastewater treatment plants and can potentially be applied wherever mechanistic models are used.

Keywords: mechanistic modelling, data-driven modelling, wastewater treatment plants

Introduction

Mechanistic models formulate empirical knowledge, such as mass balances and conversion rates, in a set of differential equations. These models can subsequently be used to simulate the behaviour of wastewater treatment plants (WWTPs), e.g. the different versions of the Activated Sludge Model (ASM) [1]. Their interpretability makes them an ideal tool for the design and analysis of WWTPs. However, extensive expert knowledge is required to design and maintain these models. This incorporated knowledge is a simplification of the plant and does not fully describe all underlying processes. For example, aeration dynamics are significantly simplified and lumped together by the $k_L a$ parameter. Complex bacterial community dynamics are crudely divided into heterotrophs and autotrophs. Mixing behaviour is completely discarded by using a continuously stirred tank reactor (CSTR) or a systemic model approach. Although these simplifications are essential for constructing an interpretable model, the trade-off is a lack of accurate predictive capabilities.

In contrast, data-driven techniques, such as linear regression and neural networks, search for relationships in available data, which is becoming more abundant. They subsequently use this gathered knowledge for simulation. Given sufficient and representative data, these models can make accurate predictions and can be automatically updated to adapt to changes in the process. However, data-driven models lack the interpretability of mechanistic models. Moreover, they are limited by the characteristics of the dataset used. They need a large amount of data representative of the whole operational space and fail to extrapolate into operational regions not seen before.

To use a model for operational purposes, such as predictive control or digital twins, one needs a model that is capable of accurate predictions, both in regimes seen and not seen before. Given that the domain knowledge incorporated is a simplification, mechanistic models do not capture all complexity and therefore do not predict accurately. On the other hand, data-driven

models fail to predict in situations not seen before. To mitigate these problems, we investigate an alternative hybrid model consisting of both regular and neural differential equations [2]. By incorporating respectively a mechanistic and data-driven technique, the strengths of both paradigms are combined. This allows for the incorporation of all available expert knowledge and data into a hybrid model. The data-driven part of the model can fill in the gaps in domain knowledge, while the mechanistic part can fill in the operational regions not represented in the dataset. Thus, this approach allows for a trade-off between data and domain knowledge, while boosting predictive accuracy.

Methods

We performed a simulation study to assess the performance of three different models:

(1) a mechanistic model using a set of ordinary differential equations of the form

$$\frac{dX}{dt} = f(X,t)\,,$$

constructed from domain knowledge, e.g. the ones present in traditional ASM models,

(2) a data-driven model using a set of neural differential equations of the form

$$\frac{dX}{dt} = nn(X,t)\,,$$

where $nn(X,t)$ is a neural network with 1 hidden layer of 100 neurons using Tanh activation functions,

(3) a hybrid model using a combination of ordinary and neural differential equations of the form

$$\frac{dX}{dt} = f(X,t) + nn(X,t).$$

The Benchmark Simulation Model 1 (BSM1) [3] is used as mechanistic (sub)models. However, to simulate imperfect domain knowledge in the mechanistic (sub)models, we adapted the BSM1 model to not account for anoxic growth of heterotrophs. We thus deliberately removed a piece of the model and thereby simplified it. Although this is an artificial construct, it allows us to study the effects of incomplete domain knowledge in mechanistic models.

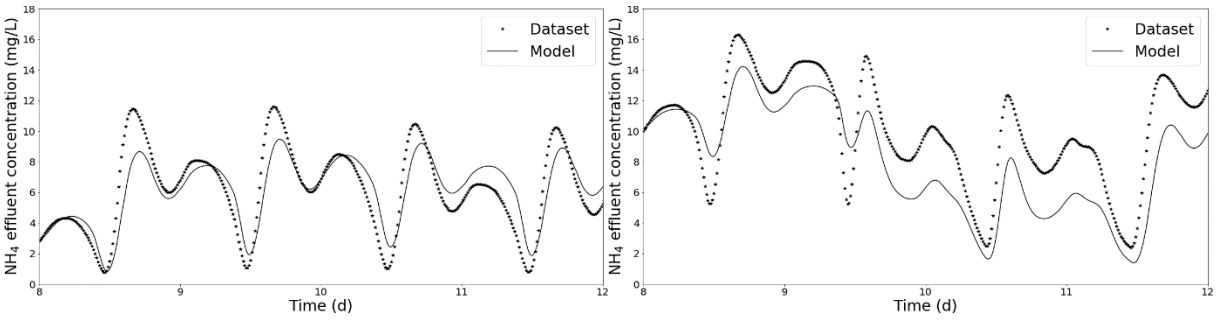
To calibrate the models, we generated synthetic data of the behaviour of a wastewater treatment plant using the full BSM1 model. The model simulated 150 days with constant influent to obtain a steady-state. Subsequently, 7 days were simulated using the first 7 days of the dry weather influent specified by BSM1. Note the use of two different versions of the BSM1 model: one version (incorporating anoxic growth of heterotrophs) to generate the dataset and one version (excluding anoxic growth) used to model this dataset.

To validate the calibrated models, we simulated 7 dry days using the last 7 days of the BSM1 dry weather influent. We subsequently used the BSM1 rainy weather influent to assess the capabilities of predicting into an operational region not represented in the original dataset. During days 8-10 of this influent dataset, a rain event occurs resulting in higher hydraulic loads and lower contaminant concentrations, an operational region highly different from dry influent.

For all simulations, we visualised the NH_4 concentration in the effluent during day 8-12, as this period includes the rain event. All simulations were performed using the PyTorch platform [4]. For each simulation, the mean root squared error (MRSE) between the ground truth is calculated.

Results and Conclusions

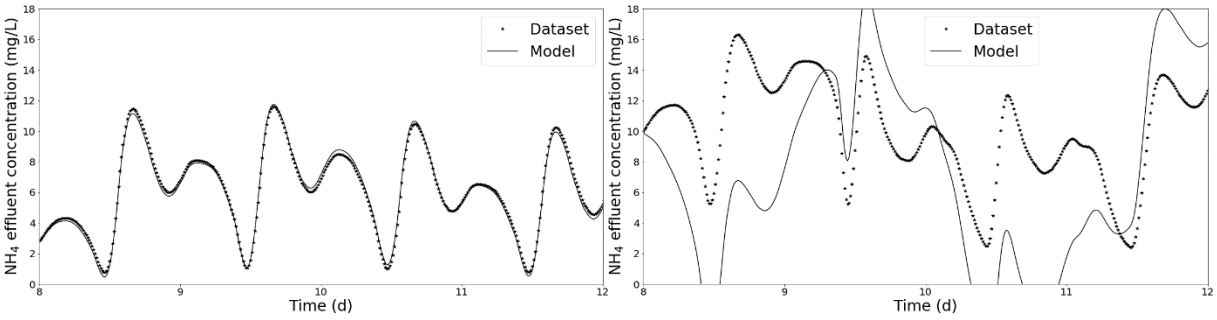
The mechanistic model is able to capture the general trends in the behaviour of the WWTP, both during dry (Fig. 1a) and rainy weather (Fig. 1b). However, as the model does not fully include all processes (i.e., it does not account for anoxic growth of heterotrophs), it is not able to accurately simulate the WWTP dynamics.



(a) Dry weather (b) Rainy weather

Figure 1. The mechanistic model, based on incomplete domain knowledge, is able to capture the general trends in the behaviour of the WWTP during both weather regimes. However, it is not able to make accurate predictions.

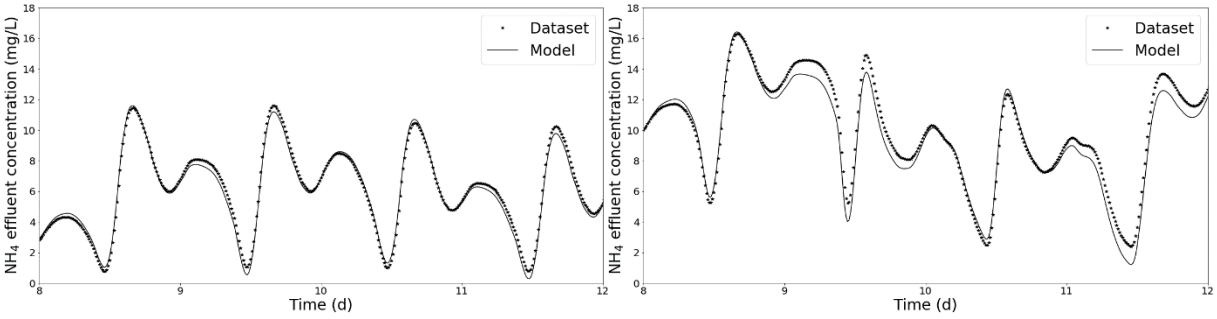
The data-driven model is able to make accurate predictions during dry weather (Fig. 2a). However, when simulating the behaviour during rainy weather (Fig. 2b), the model fails to make accurate predictions. Moreover, at times it gives physically impossible results, such as the negative NH_4 concentrations encountered. The rainy weather influent is characterized by high hydraulic loads nearly twice the size as present in the dry weather influent, and contaminant concentrations half as high. Thus, this operational regime is not represented by the original dataset and the model is in effect extrapolating, leading to poor performance.



(a) Dry weather (b) Rainy weather

Figure 2. The data-driven model is able to make accurate predictions in the dry weather regime (a). However, when a rain event occurs (b), an operational regime not represented in the original dataset, the model fails to predict accurately. Moreover, it even gives physically impossible results such as negative concentrations.

The hybrid model is able to accurately predict during dry weather (Fig. 3a). Moreover, it is able to give fairly accurate predictions during rainy weather (Fig. 3b). However, its performance is worse than during dry weather. With imperfect domain knowledge and a non-representative dataset, no submodel is able to capture the dynamics of the WWTP by itself. However, the combination of both submodels into the hybrid model allows for a better predictive performance.



(a) Dry weather (b) Rainy weather

Figure 3. The hybrid model, incorporating both an (imperfect) mechanistic and data-driven submodel, makes accurate predictions during dry weather (a). The data-driven submodel is able to fill the gaps in domain knowledge of the mechanistic submodel. However, the performance during rainy weather (b) is slightly worse, as neither perfect domain knowledge nor a representative dataset is available.

The RMSE for the different models and influent regimes is shown in Table 1. It can be seen that during dry weather, a regime represented in the dataset, the data-driven model performs the best (RMSE=5.59). The mechanistic model (39.35) is unable to fully capture the dynamics as it has imperfect domain knowledge. The hybrid model (7.62), in effect augmenting the mechanistic model with a data-driven part, is able to fill in the gaps of the imperfect mechanistic model and comes close to the performance of the purely data-driven model. When considering rainy weather, the data-driven model (149.31) performs the worst, as it is extrapolating into an unknown operational space. The mechanistic model (79.42) better captures the system dynamic but is underestimating it most of the times. The hybrid model (17.22) clearly outperforms the other models. It is able to combine the strength of both submodels and mitigates their weaknesses. However, it should be noted that a trade-off is made, as the model becomes less interpretable due to the presence of a black-box component.

Table 1. RMSE for different models and influent regimes.

Model	Dry weather (RMSE)	Rainy weather (RMSE)
Mechanistic model	39.35	79.42
Data-driven model	**5.59**	149.31
Hybrid model	7.62	**17.22**

To conclude, this simulation study shows that a hybrid model is able to combine the strengths of both the mechanistic and data-driven modelling paradigms. It allows for accurate predictions without the need for complete domain knowledge. Although the exclusion of a certain part of the domain knowledge is artificial, this study shows the potential of the hybrid approach. We are currently planning to apply this technique to a dataset of a real WWTP.

Mechanistic models are widely used in WWTPs and the wider chemical industry. The hybrid approach allows to naturally incorporate data-driven techniques into these existing models. It improves the predictive capacity, which is essential for online applications such as predictive control and digital twins. Moreover, it is possible to update these models automatically, making them adaptive to changes in the process.

References

[1] Henze, M., Gujer, W., Mino, T., & van Loosdrecht, M. C. (2000). *Activated Sludge Models ASM1, ASM2, ASM2d and ASM3*. IWA publishing.

[2] Chen, T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. In *Advances in neural information processing systems* (pp. 6571-6583).

[3] Alex, J., Benedetti, L., Copp, J., Gernaey, K. V., Jeppsson, U., Nopens, I., Pons, M.N., Steyer, J.P. & Vanrolleghem, P. (2008). Benchmark simulation model no. 1 (BSM1). *Report by the IWA Task Group on Benchmarking of Control Strategies for WWTPs*, 19-20.

[4] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017). Automatic differentiation in PyTorch.