Water Science & Technology

IWA PUBLISHING

Water Science & Technology Vol 90 No 5, 1397 doi: 10.2166/wst.2024.234

Toward smart wastewater treatment plants: a novel data-driven sludge blanket model based on stochastic differential equations

Phillip Brinck Vetter [a,*], Peter Alexander Stentoft [b], Thomas Munk-Nielsen [b], Henrik Madsen [a] and Jan Kloppenborg Møller [a]

[a] Department of Applied Mathematics and Computer Science, Section of Dynamical Systems, Technical University of Denmark, Richard Petersens Plads, Building 324, Lyngby DK-2800 Kgs, Denmark
[b] Krüger A/S, Veolia Water Technologies, Gladsaxevej 363, Søborg 2860, Denmark
*Corresponding author. E-mail: pbrve@dtu.dk

PBV, 0000-0001-9290-7942; PAS, 0000-0003-0853-3357; HM, 0000-0003-0690-3713; JKM, 0000-0002-6100-043X

ABSTRACT

A novel data-driven stochastic state space system for modeling and forecasting the sludge blanket height in secondary clarifiers is presented. The model is trained on sensor measurements of the sludge blanket height and uses as inputs (1) the clarifier sludge mass inflow rate, and (2) the clarifier recycle flow rate. The model's prediction accuracy is evaluated on data from two Danish wastewater treatment plants, for a summer and a winter month, by means of root-mean-square errors and compared with a persistence model. The model consistently outperforms the persistence model in the summer, but only one plant performs well in the winter month. The worst performing plant is challenging to model due to data quality issues and problematic (uneven and time-varying) flow distributions to the clarifiers. This led us to conclude that the best performance and stability is seen to require high data quality and well-controlled flow distribution. In summary the model achieves, in almost all cases, prediction error reductions in the order of 30–50% and 0.1–0.4 m in relative and absolute terms when compared with the predictions from a persistence model.

Key words: forecasting, parameter estimation, secondary clarifier modeling, sludge blanket height, stochastic differential equations, wastewater treatment modeling

HIGHLIGHTS

- A stochastic state space model of the sludge blanket height in secondary clarifiers at wastewater treatment plants is developed, trained on sensor measurements of the sludge blanket height and using as inputs the plant inflow rate, clarifier feed suspended solids concentration and clarifier recycle flow rate.
- The model can forecast the sludge blanket height multiple hours ahead with high accuracy, in particular on plants with good secondary clarifier flow-distribution and where high-quality sensor data is available.
- The developed model may be used in a model predictive control setup to improve secondary clarifier performance by taking future predicted behaviour into account when choosing the present control strategy for e.g. bypass fraction or recycle flow rates, which may electively increase plant hydraulic capacity, limiting severe bypass scenarios and reduce eluent concentrations.

1. INTRODUCTION

Wastewater treatment plants are a fundamental part of modern sanitation systems and serve to reduce the environmental impact on natural water bodies from human activities. The cleansing procedure at wastewater treatment plants follows a series of steps, roughly comprised of (1) mechanical treatment, (2) phase separation, and (3) biological/chemical treatment. The former step employs metal grids/screens to filter out larger objects, the second uses so-called primary clarifiers where courser materials are removed through basic settling, and the latter step employs biological tanks and secondary clarifiers to reduce various undesired chemical compounds and carbonous matters in the wastewater. This research paper focuses on the latter step, in particular on the secondary clarifier. The technique used for biological treatment is a widespread two-stage technique known as the activated sludge process, which is carried out in the connected biological basins and

secondary clarifiers. In the biological basin the wastewater is aerated in a repeated nitrification/denitrification process to increase growth rates of naturally occurring bacteria, that aid in devouring the carbonaceous matter. The wastewater is subsequently led into the secondary clarifier where the bacteria flocs, referred to as sludge, are separated from the water through gravitational settling to produce a cleansed water outflow (effluent). It is paramount that the activated sludge process is performing well, and in particular the secondary clarifier, since it is known to be the primary bottleneck of the volumes of wastewater a plant is able to handle (Stukenberg *et al.* 1983; Ekama *et al.* 1997; Ramin 2014). The secondary clarifier serves three primary purposes (Ekama *et al.* 1997): (1) producing a decontaminated effluent, (2) thickening activated sludge, and (3) storing sludge during peak loading periods. The settling dynamics of sludge in the clarifier is such that a clear sludge-water interface is established, at a particular height, across which the sludge concentration changes rapidly. This level is referred to as the sludge blanket height. A high sludge blanket is correlated with high concentration of bacteria (called suspended solids) in the effluent (Ekama *et al.* 1997) why it is of interest to keep it at a sufficiently low level. A fundamental challenge for treatment plants is that the designed hydraulic capacity i.e. the maximum permissible flow rate through the plant, is often surpassed during events of heavy precipitation, compromising the activated sludge operation. In a worst-case scenario it can be necessary to skip the activated sludge process altogether, a process known as *bypass*, to prevent diluting the sludge mass, or to avoid increased effluent concentrations as a result of a high sludge blanket. It is financially infeasible to overcome this challenge by building larger treatment plants, because the required volumes are immense. In the pursuit of limiting human environmental impact treatment plants are under increasingly stringent demands to further reduce effluent contamination and lower bypass volumes. It is thus crucial to continue to develop and improve the operation and control of treatment plants, and in particular the activated sludge operation, in order to better exploit plant hydraulic capacities and reduce effluent contamination.

There has been substantial research into clarifier behavior and settling dynamics and we provide a brief overview of the history in the following. The earliest research by Coe & Clevenger (1916) discussed solids capacities in the layers of the clarifier, but the sedimentation theory was not thoroughly discussed until Anderson & Gould (1945), with the introduction of the fundamental solid-flux theory in Kynch (1952), and the further elaboration of so-called hindered settling by Vesilind (1974). Towards the 1990s research focused on improving clarifier design based on the theory (Fitch 1966; Kos 1977; Tekippe & Bender 1987), and an operational so-called *state point* strategy for dealing with overloaded clarifiers were demonstrated in Keinath *et al.* (1977) and Keinath (1985, 1990). The state point strategy lead to the first online control strategies known as step feed to prevent bypass during wet weather conditions (Thompson *et al.* 1989). In combination with a rule-based four step control system (Balslev *et al.* 1994) developed an online control strategy for the secondary clarifier, which laid the foundations for the clarifier control implemented at various Danish treatment plants (due to Krüger Veolia's digital software tool *STAR Utility Solutions*™). One-dimensional concentration models based on the sedimentation theory of Kynch (1952) were addressed extensively in the 1990s by Takács *et al.* (1991), Hamilton *et al.* (1992), and Watts *et al.* (1996) and others, but these models were primarily adapted using steady-state data, although a dynamic model was introduced by Clarcq *et al.* (2003). The Takács *et al.* (1991) models are obtained from a spatial finite difference scheme of the advection-diffusion partial differential equation, with incorporation of settling dynamics from empirical flux settling curves and settling parameters. These models continue to form the basis of one-dimensional clarifier modeling, and the more recent work in the literature improves on the mathematical descriptions of, e.g., mechanistic compression dynamics and solids dispersion, and generally focuses on appropriate mathematical formulations to ensure well posedness and stability. The reader should consult the work of, e.g., Plosz *et al.* (2011), Bürger *et al.* (2012), and Guyonvarch *et al.* (2015) for a solid overview for these modern formulations. The newer contributions in the literature also considers uncertainty quantification (Guyonvarch *et al.* 2020; Zhou & Li 2023) and the effect of certain filamentous bacteria on settling properties (Qiu *et al.* 2023). The literature also contains many three-dimensional computational fluid dynamics models, e.g., Xu *et al.* (2022), that are used for optimizing design and similar research. The reason that these models are not used here, is due to their large computational complexity, making them unviable for online operations that require fast continuous calculations.

The general challenges with respect to sludge blanket height prediction is firstly that clarifier behavior in operational settings is very complicated, and one-dimensional models must (obviously) neglect more complicated features such as, e.g. turbulence, geometric properties, or uneven sludge distributions, and consequently suffer accuracy impairments. A primary difficulty for the models found in the literature is furthermore that inferring the sludge blanket height ultimately comes down to inferring a single point on the estimated sludge concentration profile curve. The models are also difficult

to calibrate because continuous concentration profile measurements are generally not available. The models are relatively large with approximate 10–100 states, so optimization across multiple weeks of training data is computationally quite expensive. Similarly expensive is forecasting, which is essential for model predictive control where the control signal(s) must be repeatedly optimized over. To overcome these prediction challenges, we present a novel data-driven model which specifically targets the sludge blanket height directly, as opposed to indirectly through concentrations. A primary advantage here is that sludge blanket measurements are used directly to calibrate the model. The model furthermore only consists of a single state equation, and is thus fast and easy to evaluate in online settings making it ideal in a model predictive control setup. A challenge with the proposed model, as opposed to the existing models, is that it lacks clear physical interpretation, being based on a very simple qualitative understanding of the clarifier dynamics, which inevitably neglects more complicated dynamics. Due to the models simplicity high quality data is crucial to be able to identify appropriate correlations between the exogenous model variables (plant inflow, clarifier suspended solid concentrations, clarifier recycle flow rate) and the endogenous sludge blanket height. A challenge in this regard is that there is a very limited understanding of the quality of the blanket measurements and their representativeness. The measurements are conducted only at one particular location, and thus wrongly assumes a uniform blanket height, and accuracy impairments due to, e.g., turbulence in the water column is unknown. There is currently, to the best of the authors' knowledge, no similar model in the literature, so no direct comparisons can be made.

The model is a one-dimensional continuous-discrete state space system (Jazwinski 1970) consisting of a stochastic differential equation (SDE) and an algebraic observation equation (AOE). We refer to the model as belonging to the class of *gray-box* models, combining black-box statistical models with traditional 'white-box' differential equation models. A stochastic differential equation can be viewed as a generalization of an ordinary differential equation with a *natural* incorporation of uncertainty through the addition of a so-called diffusion term whose increments are assigned a probability distribution. The solution to a stochastic differential equation at every point in time is therefore a probability distribution, rather than a single point. This allows for much more model flexibility when fitting to data where the model can adapt locally to observations, and that results in parameter estimates that better reflect the true system. This is in contrast to ordinary differential equations where local changes to the solution curve, to better match data, will have global implications (Møller 2011). In essence the role of the stochastic differential equation is to transform the data-fitting procedure into a weighted fit based on the uncertainty (the probability distribution) assigned by the stochastic differential equation, and the uncertainty of the observations. The developed model is fitted to data by estimating the parameters that minimize the negative log-likelihood of independent one-step state transition probabilities, where state filtration is carried out using an Extended Kalman Filter. The procedure takes advantage of algorithmic differentiation and the speed of **C++**, which is made available through the **R** *package* **TMB** (Kristensen *et al.* 2016). The work presented is inspired by the work in Stentoft *et al.* (2019, 2021) where a similar model for the biology was developed for forecasting and controlling the aeration process in the nitrification/denitrification step of the biological treatment basins. The authors also demonstrated the ability to implement various model predictive control strategies tailored towards, e.g. cost-savings, or reduced carbon emissions by considering electricity prices. The overarching goal here is to construct and link together similar model predictive control frameworks for various units at the treatment plant to increase performance, but also to embed flexibility in the sense of Junker *et al.* (2020) into treatment plants such that the operation can respond to variations in energy prices.

The paper is organized as follows: In Section 2 we present the two wastewater treatment plants whose data the present analysis is based on, and discuss properties of this data. In addition to that we briefly discuss theory of stochastic state space systems, likelihood estimation, the developed model, theoretical details, and numerical settings. In Section 3, we present estimation results, model residuals, showcase example predictions, prediction accuracy in terms of root-mean-square error , and finally consider predictions with uncertain (forecasted) inputs. The section is concluded with a discussion of the results, the chosen methods and computation techniques, subtle challenges with regards to model training and more. Finally, a conclusion is given in Section 4.

2. METHODS

The results presented in this article are computed using the statistical software language **R** (R Core Team 2023), the plots/graphics are created using the **ggplot2** package (Wickham 2016), tables are created using the **knitr** (Xie 2014) and **kableExtra** (Zhu 2021) packages, and likelihood calculations are based on the **TMB** package (Kristensen *et al.* 2016).

2.1. Model implementation and workflow

The model is implemented as follows: First, a likelihood function for the observed data, based on the Extended Kalman Filter algorithm similar to that discussed in Brok *et al.* (2018) is implemented in **C++**. The **C++** script uses among other things a set of macros made available by **TMB**. The reader is referred to **TMBs** GitHub Pages documentation for further information about implementation procedures. Once this is done the likelihood function, and its gradient and hessian, is made available by TMB for use in R. This allows for straight-forward minimization of the likelihood function and thus obtaining the best fitting parameter and the associated states. Once the parameters are obtained predictions can be made by computing prior mean and variance state estimates through the Extended Kalman Filter algorithm, which amounts to solving a set of moment ODEs (17). This entire procedure is made easily available through the authors' own (under-development) package **ctsmTMB**, which has been used to generate the results presented in this article, and the reader is referred to the package's GitHub Page for more information. In addition an example script is also provided here which demonstrates how estimation and prediction is carried out.

2.2. Treatment plants and data

The data used in this study was obtained from two wastewater treatment plants in Denmark, *'Damhusåen'* and *'Kolding Central'*. Damhusåen is one of three larger treatment plants that serve the Copenhagen metropolitan area. The facility receives wastewater from 15 municipalities and leads its effluent into *Øresund*, the strait separating Denmark and Sweden. Kolding Central near Agtrup is located in Southern Jutland. It is the largest treatment plant in the municipality of Kolding and leads its effluent into *Little Belt*, the strait between the island of Funen and Jutland in Denmark. The two treatment plants both employ mechanical, biological and chemical treatment. The former is comprised of four activated sludge operation lines each consisting of a biological tank pair and six rectangular secondary clarifiers, while the latter consists of a single operation line with four circular secondary clarifiers. A summary overview of the plants' statistics are shown in Table 1, and a layout is presented in Figure 1 with colored circular marks indicating the presence of available measurements, which comprise (1) the sludge blanket height h_t in each secondary clarifier, (2) the plant inflow rate $Q_{f,t}$, (3) the suspended solids concentration in each pair of biological tanks $C_{f,t}$ and (4) the recycle flow rate out of each clarifier $Q_{r,t}$. The subscripts f and r in this notation stem from *feed* and *recycle*, respectively, and t is a time index adopted from the mathematical literature on stochastic processes. The attentive reader will notice the '?' markings at the flow distribution channels at Damhusåen which emphasize the challenging property that the plant inflow is unevenly distributed between both operation lines and clarifiers in a non-uniform and time-varying way.

The model development presented in this article has been based on 6 months of data collected from July to December in 2,019 and 2022, at Damhusåen and Kolding Central, respectively. The data were acquired from the two treatment plants through Krüger Veolia's Hubgrade™ cloud solution. The sludge blanket measurements are collected by an ultra-sonic sensor (Hach 2023) installed above the surface of the clarifiers, and calculated based on the return time of an emitted ultra-sound signal. We use these calculated values directly as the sludge blanket height, without any further processing. Each data signal is sampled roughly every 2 min but in an unsynchronized fashion, so joint time-stamps for all signals are achieved by linear interpolation. The data was subsequently aggregated into 10 min intervals, using median values, to ease the computational effort. A few example tests demonstrated that the results (i.e. the model parameter estimates) were unaffected by this aggregation, arguably since blanket dynamics are much slower. The two selected periods were chosen because the data contained limited signal artifacts such as drop-outs, unrealistic jumps or peculiar oscillations. In this article, however, we focus just on August and December to exemplify the model's properties and predictive performance. The purpose of selecting these two months is to showcase performance differences caused by seasonality i.e. 'summer' and 'winter', and the reason for comparing these two treatment plants is due to a clear difference in the plant operating conditions. There are four operating lines at

Table 1 | The size of the two included treatment plants which includes their population equivalent (PE) capacities, the number of operation lines, the number of secondary clarifiers, and the dimensions of the clarifiers

	PE capacity	Lines	Clarifiers	Shape	Dimensions (meters)
Damhusåen	350.000	4	24	Rectangular (H×L×D)	5×20×3
Kolding Central	125.000	1	4	Circular (Dia×D)	30×4

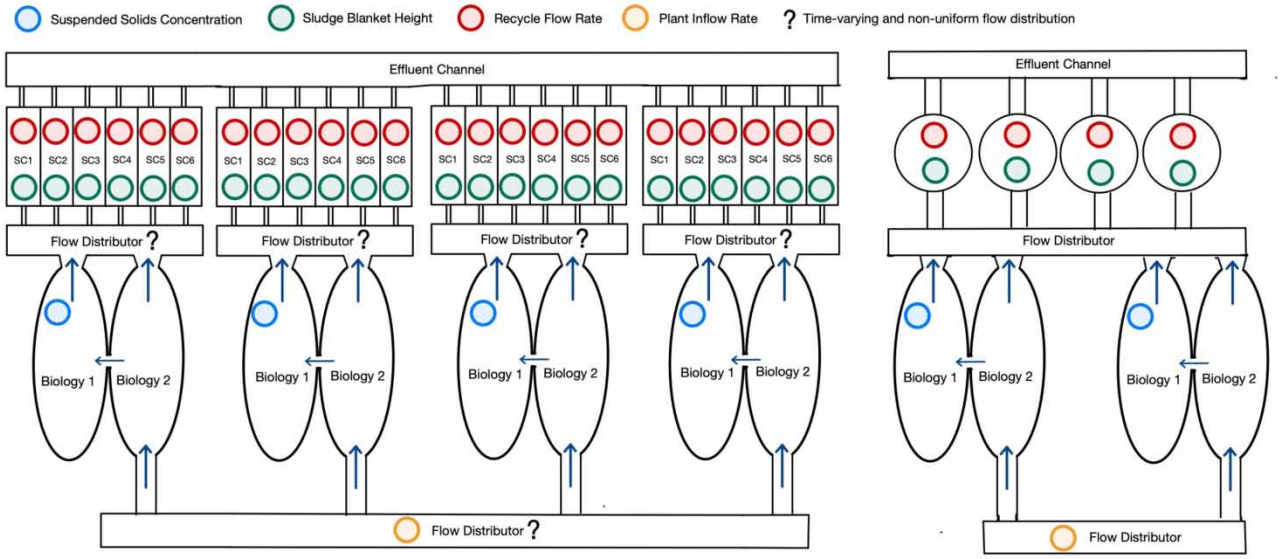

Figure 1 | A sketch illustrating the layout of Damhusåen, wastewater treatment plant (WWTP) (left) and Kolding Central WWTP (right), marked with circles of various colors corresponding to the location of sensors.

Damhusåen but we only consider one of the lines in this article for sake of simplicity. The second operation line was deliberately not picked since it stood out as having particularly large dispersion, but we picked arbitrarily among the remaining three lines, so results are most likely representative of the general behavior. The reader may appreciate the dispersion differences by comparing the chosen data shown in Figure 2 with the discarded line placed in Figure 9 of Appendix A. The figures contain time series data of the sludge blanket heights h_t for each operation line together with the (normalized) mass inflow rates $M_{f,t} = Q_{f,t}C_{f,t}$. The measured sludge blanket heights are shown as gray-colored regions spanning minimum to maximum, but the individual sludge blanket time series are omitted. The intention here is to emphasize the blanket dispersion between lines and plants, which evidently is much greater at Damhusåen than at Kolding Central. The gray-colored regions in Figure 2(c) and 2(d) are very thin indicating close-to-identical behavior between all clarifiers. In contrast these regions are seen to be much larger at Damhusåen in particular in Figure 2(b). The mass inflow rates are included in the figures to showcase its correlation to the sludge blanket height. The recycle flow rates are omitted to avoid cluttering. The plant inflow rate and suspended solids concentrations are also omitted in favor of the mass inflow rate, which we expect sums up their contributions. That being said there may be more complicated interactions from the terms individually that we suppress in favor of a clearer interpretation. Generally we expect the sludge blanket height to mirror the behavior of the mass inflow rate. This expectation is generally met in the presented data, but there are evidently also some more complicated behavior present in many situations. In the following we highlight some of the scenarios where the sludge blanket response is interesting, or unexpected, in order to emphasize some of the challenges in the data.

Damhusåen:

(1) **August 2–8 and 21–31:** The sludge blankets are much higher in the former period than the latter for similar mass inflow rates
(2) **August 8–23:** The sludge blankets and mass inflow peaks coincide on the 16th, 18th, 19th and 21st, but are delayed in time on the 13th–14th and 17th–20th.
(3) **August 11–13:** The sludge blankets are much higher at the end of the month, even though the mass inflow rate is similar.
(4) **August 28–29:** The sludge blankets do not respond to the small peak in the mass inflow rate.
(5) **December 1–31:** The sludge blankets are gradually increasing throughout the month although the non-peak mass inflow rate remains roughly the same.
(6) **December 15–16:** The sludge blankets do not respond to the large increase in the mass inflow rate.
(7) **December 18–20:** The sludge blanket height reductions here are (very) large when compared to the complete lack of blanket response on the 21st–22nd where a similar low mass inflow rate is obtained (although only for a short duration).

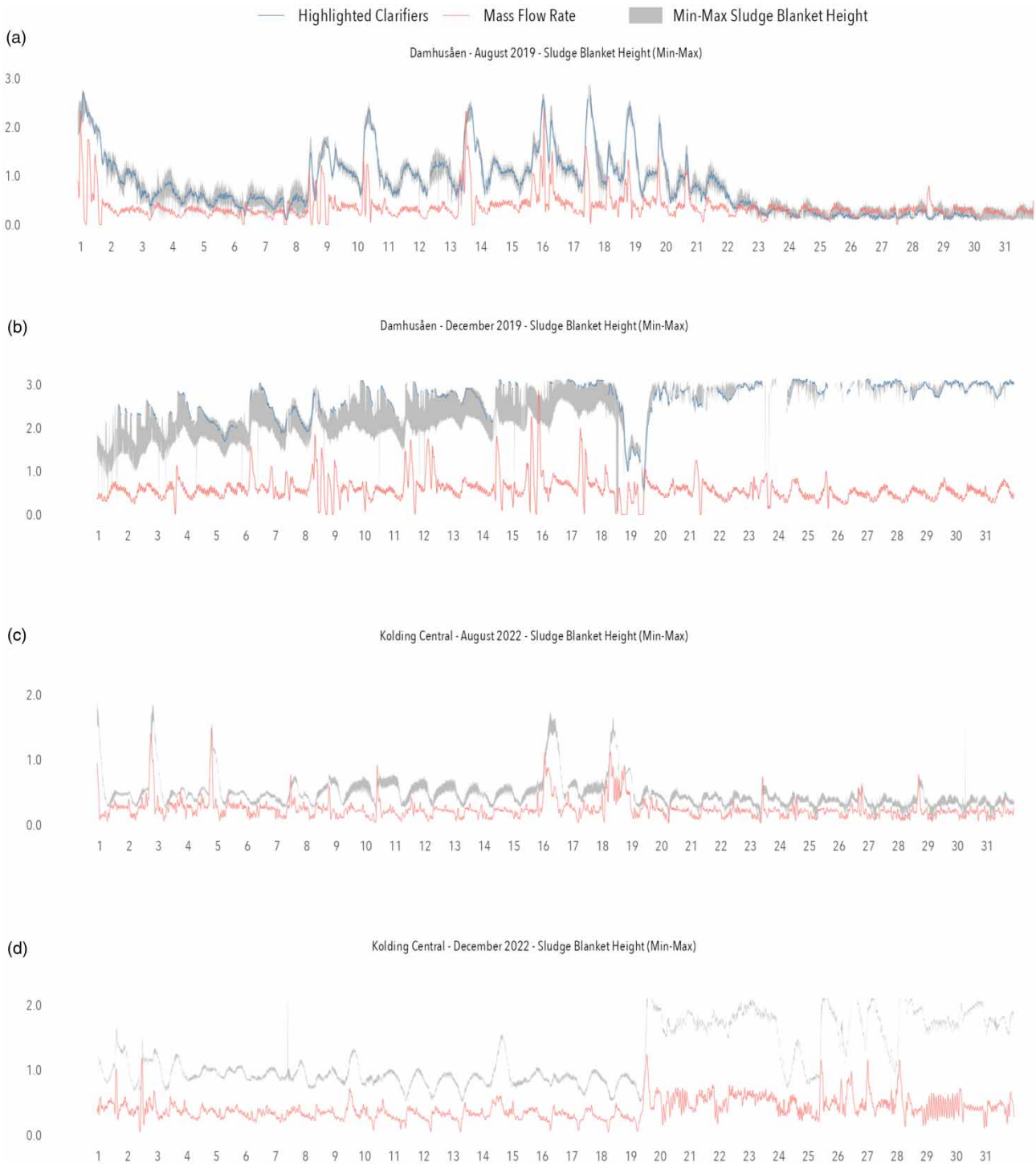

Figure 2 | The figure shows the data analyzed in this article from (**a**) and (**b**) Damhusåen, (August and December 2019) and (**b**) and (**c**) Kolding Central (August and December 2022). The time series input shown is the mass inflow rate. The individual blanket lines are not shown, rather the maximum and minimum blanket height in a particular line is shown as a gray-colored area.

Kolding Central:

(1) **August 5–7 and 11–15:** The blanket heights are lower in the former period when compared to the latter even though the mass inflow rate is comparable.

(2) **August 1–15:** The diurnal blanket oscillations reach a new level from the 8th and onwards even though the mass inflow rate appears unchanged during the period.
(3) **August 3, 5, 16–17 and 28:** The blanket peaks that occurs on the 16th–17th exhibit a larger time-delay than on the other days relative to the mass inflow rate peak.
(4) **December 2–3 vs 19–20 and 9–10 vs 14–15:** The blanket responses on the latter dates are larger than the responses on the former even though the mass inflow rates are similar.
(5) **December 19–31:** The blanket dynamics seem to change abruptly after the 19th, and the sensitivity to the mass inflow rate appears to increase. In particular, the sludge blanket response to the mass inflow rate is much greater on the 29th–30th than earlier in the month.
(6) **December 24:** The blanket heights decreased substantially even though the change in the mass inflow rate was minor.

These examples of more complex blanket behavior may arise from nonlinear effects such as hysteresis, the mentioned non-uniform clarifier flow distribution or other excluded system drivers such as bacteria filament types and clarifier geometry . The model that we present shortly will generally have difficulty capturing these nonlinear behaviors, at least if the system undergoes a relatively fast transition into a state where the dynamics are significant different. Conversely, however, if the dynamics change slowly across time i.e. in a weekly, monthly or quarterly fashion then we expect that the 'new' dynamics may be learned through re-estimation of the model parameters, as noted by Bürger *et al*. (2011).

2.3. Theory

The model we propose here is in the form of a continuous-discrete stochastic state space system. This is a model which is comprised of (1) a stochastic differential equation that describes the evolution of (the probability distribution of) a continuous time state x_t and (2) an observation equation that links the state to some observations made at discrete points in time with some assumed noise. Mathematically we write this as:

$$\mathrm{d}x_t = f(x_t, u_t, \theta)\,\mathrm{d}t + g(x_t, u_t, \theta)\,\mathrm{d}\omega(t) \tag{1}$$

$$z_k = h(x_{t_k}, u_{t_k}, \theta) + S_k(x_{t_k}, u_{t_k}, \theta)\varepsilon_k, \tag{2}$$

with state(s) $x \in \mathbb{R}^{n_s}$, input(s) $u \in \mathbb{R}^{n_u}$, parameter(s) $\theta \in \mathbb{R}^{n_\theta}$, observations $z \in \mathbb{R}^{n_m}$, incremental Wiener processes $\mathrm{d}\omega \in \mathbb{R}^{n_\omega}$, drift function $f \in \mathbb{R}^{n_s}$, diffusion function $g \in \mathbb{R}^{n_s} \times \mathbb{R}^{n_\omega}$, and observation function $h \in \mathbb{R}^{n_m}$. The quantity ε_k is a probability density that quantifies the noise associated with the measurement procedure. In the common case where we assume that the noise is standard normal i.e. $\varepsilon_k \sim \mathcal{N}(0, 1)$, then it directly determines the conditional distribution $z_k|x_{t_k} \sim \mathcal{N}(h(x_{t_k}, u_{t_k}, \theta), S_k^2(x_{t_k}, u_{t_k}, \theta))$. We sought an estimate of the model parameters θ that based on the observed information $\mathcal{G}_N = \{z_0, z_1, , z_{N-1}, z_N\}$ minimizes the associated negative log-likelihood function. The negative log-likelihood is the joint probability density of all the observations viewed as a function of the parameter θ, so the minimizer is given by:

$$\hat{\theta} = \underset{\theta}{argmin}\{-\log L(\theta; \mathcal{G}_N)\} \tag{3}$$

$$= \underset{\theta}{argmin}\{-\log \mathcal{F}_{\mathcal{G}_N}(z_0, z_1, \ldots, z_N; \theta)\} \tag{4}$$

$$= \underset{\theta}{argmin}\left\{-\sum_{i=1}^{N} \log \mathcal{F}_{Z_k|\mathcal{G}_{k-1}}(z_k \mid z_0, \ldots z_{k-1}; \theta)\right\}, \tag{5}$$

where $\mathcal{F}$ are appropriate densities. A standard approach to obtaining (5) is to invoke the law of total probability on each of these N terms and condition on the state X_{t_k}. This yields individual terms:

$$\mathcal{F}_{Z_k|\mathcal{G}_{k-1}}(z_k \mid z_0, \ldots z_{k-1}; \theta) \tag{6}$$

$$= \int_X \underbrace{\mathcal{F}_{Z_k|X_{t_k}}(z_k \mid x; \theta)}_{\text{likelihood}} \underbrace{\mathcal{F}_{X_{t_k}|\mathcal{G}_{k-1}}(x \mid z_0, \ldots z_{k-1}; \theta)}_{\text{prior}}\,\mathrm{d}x, \tag{7}$$

where the first term in the integral has no longer a conditioning on the past information $\mathcal{G}_{k-1}$ because that information is by the Markov property contained in X_{t_k}. This form is convenient because it involves the following two terms (1) the likelihood distribution partially provided by ε_k, as mentioned above and (2) the (prior) distribution of the state (at time t_k) given all previous observations (up to time t_{k-1}). The former is known once ε_k has been provided. The latter is obtained by integrating the stochastic differential equation forward in time, which generally involves solving the Fokker–Plank equation associated with the stochastic differential equation. This is usually intractable and computationally expensive so in practice approximations are sought. In this work we employ the continuous-discrete time Extended Kalman filter, whose Gaussian assumptions on ε and on the state distribution of X_t allows for an analytical evaluation of the integral in (7), such that (5) becomes a sum of normal distributions. The reader is referred to Brok *et al.* (2018) for a quick walk-through of the Extended Kalman Filter and to Madsen *et al.* (2015) and Thygesen (2023) for more elaborate discussions on filtering principles and likelihood theory.

Now, the actual proposed model here is an augmentation of the so-called Ornstein-Uhlenbeck process (8). The model describes a system that is attracted towards a (mean) value μ, with an exponential decay with a characteristic time of α. We argue that such a simple system roughly describes the qualitative dynamics of the sludge blanket height. This should be understood in the sense that for fixed operating conditions a unique concentration profile is obtained in the clarifier (Jeppsson & Diehl 1996), and thus a stationary blanket level is too, and it is this level that we interpret as μ. We may further interpret α as containing information about the settling characteristics of the sludge, which in turn affects the velocity with which the blanket level changes. A simple low-dimensional model like this is advantageous partly due to its linear properties, which promotes computational speed and robustness. Since we have access to sludge blanket sensors, the observation equation is just the identity function (9), and the state space system is therefore given by:

$$\mathrm{d}h_t = \frac{1}{\alpha}(\mu - h_t)\,\mathrm{d}t + \sigma\,\mathrm{d}\omega_t \tag{8}$$

$$z_{t_k} = h_{t_k} + \varepsilon_{t_k}, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_z^2). \tag{9}$$

To reiterate; the latent state h_t here is the true but unknown sludge blanket height at time t, which is directly observed as z_{t_k}, but subject to normally distributed noise e_{t_k}. The conditional mean of (8) is given by:

$$\mathbb{E}(h_t|\mathbb{E}(h_0) = \mu_0) = \mu + (\mu_0 - \mu)\mathrm{e}^{-\alpha^{-1}t} \to \mu \quad \text{for} \quad t >> \alpha \tag{10}$$

and time to reach 95% of this new level is roughly $T_s = 3\alpha$ since $\exp(-\alpha^{-1}3\alpha) = \exp(-3) \approx 0.05$. The random fluctuations imparted by the diffusion $\sigma\,\mathrm{d}\omega_t$ carries into the conditional variance:

$$\mathbb{V}(h_t|\mathbb{V}(h_0) = \sigma_0^2) = \frac{\sigma^2\alpha}{2}(1 - \mathrm{e}^{-2\alpha^{-1}t}) + \frac{\sigma_0^2\alpha}{2}\mathrm{e}^{-2\alpha^{-1}t} \to \frac{\sigma^2\alpha}{2} \quad \text{for } t >> \alpha. \tag{11}$$

For further theoretical properties of the Ornstein–Uhlenbeck process the reader is referred to Iacus (2008). Now, in order to capture variations in the steady-state value when the operating conditions change we introduce a time-dependency to the drift parameters $\{\alpha, \mu\}$ by expanding them as functions of the inputs $u_t = \{M_{f,t}, Q_{r,t}, Q_{f,t}, C_{f,t}\}$, where we note that the plant inflow rate and suspended solids concentrations are included as individual terms. The functional relationships between $\{\alpha, \mu\}$ and U_t were initially investigated by augmenting individual parameters in the system as random walks, one at a time, inspired by the work of Møller *et al.* (2012) and as carried out in Vetter (2020). The general approach is to seek parametric approximations between the smoothed state estimates and input(s), by inspection of the point cloud that emerges when plotting these against one another. This approach did however this not yield any meaningful relationships. As an alternative strategy we resorted to a brute-force monomial expansion in U_t up to second order, e.g.

$$\mu_t = b_0 + b_1M_{f,t} + b_2Q_{r,t} + b_3Q_{f,t} + b_4C_{f,t} + b_5M_{f,t}^2 + b_6Q_{r,t}^2 + b_7Q_{f,t}^2 + b_8C_{f,t}^2. \tag{12}$$

The optimal expansions were found to be:

$$\alpha_t = \alpha_0, \tag{13}$$

$$\mu_t = b_0 + b_1 M_{f,t} + b_2 Q_{r,t}. \tag{14}$$

The meaning of optimal here should be understood as yielding the best predictive performance. Specifically we compared the various models' predictive performance, in terms of root-mean square errors (RMSEs), across all 6 months, using estimated parameter values from the previous month's data. The model in (15) outperformed or was on par with all other formulations, and was furthermore significantly faster and more robust to optimize. The robustness of the other models were presumably impaired due to (1) difficulties with optimizer convergence in the larger parameter space and (2) a sensitivity to input-dependence in α, which can cause the numerical solution of the mean and variance ODEs to fail. This latter fact may explain why the zeroth order expansion in (14) was preferred. The expansion found in (15) reveals that the two most important terms for accurately predicting the sludge blanket level is the sludge mass inflow rate and the recycle flow rate. Intuitively we expect the former to account for an increasing sludge blanket through the addition of more sludge into the clarifier, while the latter is expected to decrease the sludge blanket through the removal of sludge from the bottom of the clarifier.

2.4. Optimizer and parameter settings

We impose two parameter transformations to ensure robustness and appropriate parameter domains. The two transformations are (1) an inverse logit transformation on μ_t such that $\mu_t \in (0, H)$ (i.e. within the clarifier), and (2) an exponential transform to α and σ to ensure positivity. The final form of the presented stochastic differential equation therefore becomes:

$$\mathrm{d}h_t = \mathrm{e}^{-\tilde{\alpha}}(H\ invlogit\ (b_0 + b_1 M_{f,t} + b_2 Q_{r,t}) - h_t)\,\mathrm{d}t + \mathrm{e}^{\tilde{\sigma}}\,\mathrm{d}\omega_t, \tag{15}$$

where $\alpha_0 = \exp(\tilde{\alpha})$ and $\sigma = \exp(\tilde{\sigma})$. The inverse logit transformation on μ_t is particularly important when the model is used to forecast based on input values outside the range of the training data, where the fitted monomial might extrapolate poorly.

The parameter estimation was performed using the **nlminb** optimization algorithm (Gay 1990) available from the **stats** package (R-Stats-Documentation 2023). The parameters to optimize are $\theta = \{\tilde{\alpha}, b_0, b_1, b_2, \tilde{\sigma}\}$ with fixed observation noise $\sigma_z = 0.05/0.10$ at Kolding Central and Damhusåen, respectively. While the observation noise parameter can be estimated too, it is often helpful to reduce the number of noise parameters since these can be difficult to identify separately. The chosen value for σ_z was based on the reported accuracy level of a commercially available sludge blanket detector (Hach 2023). This should be interpreted in terms of the associated 95% confidence interval i.e. the sludge blanket detector has an accuracy of $\pm 2\sigma_z$. The optimization was initialized with the parameter values $\theta_0 = \{0, 10^{-5}, 10^{-5}, 10^{-5}, \log 10^{-2}\}$. The lower and upper parameter boundaries were set at $\tilde{\alpha} \in [1/6, 24]$, somewhat arbitrary bounds for $\tilde{\sigma} \in [\log 10^{-10}, 0]$ and with b_i's unconstrained. We noted that optimizer stability required low initial values of b_i's, otherwise the optimizer explodes. A district advantage of the **nlminb** optimizer is its possibility of using the objective function hessian during optimization. This is particularly ideal here because **TMB** also returns the hessian through algorithmic differentiation. The hessian-aided optimization was found to be superior to the gradient-only optimization insofar as it was actually able to converge to a true minimizer. The gradient-based search although faster was only able to locate the same minimizer as the hessian-aided optimization for one particular month. In that case (November, $\approx$ 4,000 observations) the computation time using the hessian increased to ≈ 2 seconds up from ≈ 1.2 seconds using only the gradient. As an example of the behavior for the other months we highlight the month of August of Figure 2: In this case, the hessian-aided and gradient-only optimizations required ≈ 2.0 *and* 0.2 seconds, respectively, but although **nlminb** reported convergence in both cases (*both X-convergence and relative convergence*), a closer inspection on the maximum gradient components revealed $4.7 \cdot 10^{-8}$ and $5.9 \cdot 10^{1}$, respectively, proving that the gradient-only minimizer was false. The negative log-likelihood difference was also substantial for the two with ≈ -7758 and ≈ -7133, respectively. A similar pattern was seen in the other 5 months in the available data set.

3. RESULTS AND DISCUSSION

The following section presents an analysis of the modeling results and includes estimated parameter values, example predictions, and an evaluation of the model performance. In order to avoid an excessive number of figures in the main article the reader is referred to Appendix B for the results from Damhusåen. The parameter estimates and example predictions are made

based on the data from a single selected clarifier to exemplify the model properties, since it is impractical to include results for all (4 or 6) available clarifiers. The results shown are representative of all clarifiers on Kolding Central, but this is not the case for Damhusåen. The reason that Damhusåen results are not representative is due to the large blanket variation observed there, as shown in the data in Section 2.2, which is believed to be caused by the plant's flow distribution characteristics. We will discuss this further in Section 3.3.

3.1. Model estimation and validation

We present parameter estimates and associated statistics in Tables 2 and 4, but emphasize again that the parameter estimates at Damhusåen are not representative of all clarifiers. To emphasize this point consider the range of values obtained from estimating on all Damhusåen clarifiers for August; $\alpha \in [2.06, 2.48]$, $b_1 \in [1.99, 2.85]$ and $b_2 \in [-0.71, -0.31]$, and for December; $\alpha \in [1.62, 3.18]$, $b_1 \in [1.63, 8.37]$ and $b_2 \in [-0.87, -2.23]$. We note in particular that for December the values of b_1 and b_2 span large ranges, indicating different sensitivity to changes in the mass inflow rate, and the recycle flow rate, respectively. The differences in the time-constant α is also large ($\exp(1.62) \approx 5$ versus $\exp(3.18) \approx 24$), implying that the sludge blanket responds much slower/faster in certain clarifiers. Returning to inspecting the parameter estimates in the two tables, we draw the following conclusions:

(1) The estimated time-constants for the two plants are found to be in the order of $\alpha = \exp(\tilde{\alpha}) \approx 5 - 8$ with mean values across all clarifiers of $\alpha \approx 5.75$ h and $\alpha \approx 9$ h at Kolding Central and Damhusåen, respectively. According to the model the stationary sludge blanket height determined by μ_t is thus reached to 95% of its value in $3\alpha \approx 17 - 27$ h under fixed inputs. Judging from various peaks in the considered data (Figure 2) the estimate for Kolding Central seems appropriate, while it is more difficult to tell from Damhusåen where certain situations (e.g., after the peaks on the 1st, and the 20th of August) indicate that such a process takes closer to 36–48 h instead.
(2) The estimates of b_1 and b_2 carry the expected signs (positive and negative, respectively), indicating that if the mass inflow rate increases then so does the (stationary) sludge blanket height, and vice versa for the recycle flow rate.
(3) The larger dispersion at Damhusåen is seen to result in process noise parameters that are $\exp(1) \approx 3$ times larger than those at Kolding Central, an increase from approximately $\exp(-3)$ up to $\exp(-2)$.

The estimated uncertainties are relatively small, as evident from the provided t-test statistics, and all parameters are thus significant. It should be noted that the null hypotheses (being different from zero) are only relevant for b_1 and b_2 since the remaining parameters are both transformed and not related to any of the inputs. The sludge blanket sensitivity to the inputs can be compared using the provided *scaled estimates* – these are simply b_1 and b_2 scaled by the median input value of $M_{f,t}$ and $Q_{r,t}$, respectively. First, inspecting the values from Kolding Central reveal that the relative impact of the two inputs changes in favor of the mass inflow rate from roughly 1:1 to 6:1 between August and December. The converse conclusion is found on Damhusåen with a decrease from 6:1 down to 1:1 based on Table 4, but the average effect for all six clarifiers showed a decrease from 5:1 down to 3:1. This suggests that the recycle flow rate becomes less impactful in terms of controlling the sludge blanket height on Kolding Central during winter, but becomes more impactful on Damhusåen. We generally

Table 2 | Estimated parameter values, associated standard errors, t-test statistic and input-scaled (using median input values) parameter estimates at Kolding Central for the data shown in Figure 2(c) and 2(d), for a single clarifier

August					December				
	Estimate	**Std. error**	***t*-test**	**Scaled estimate**		**Estimate**	**Std. error**	***t*-test**	**Scaled estimate**
$\tilde{\alpha}$	1.596502	0.031724	50		$\tilde{\alpha}$	2.101859	0.034789	60	
b_0	−2.229816	0.060314	37		b_0	−3.095960	0.110200	28	
b_1	0.000154	0.000005	33	0.883888	b_1	0.000275	0.000012	23	2.757868
b_2	−0.009058	0.000818	11	−0.711820	b_2	−0.003085	0.000280	11	−0.459864
$\tilde{\sigma}_x$	−3.357198	0.031832	105		$\tilde{\sigma}_x$	−2.972295	0.022831	130	
σ_y	0.050000				σ_y	0.050000			

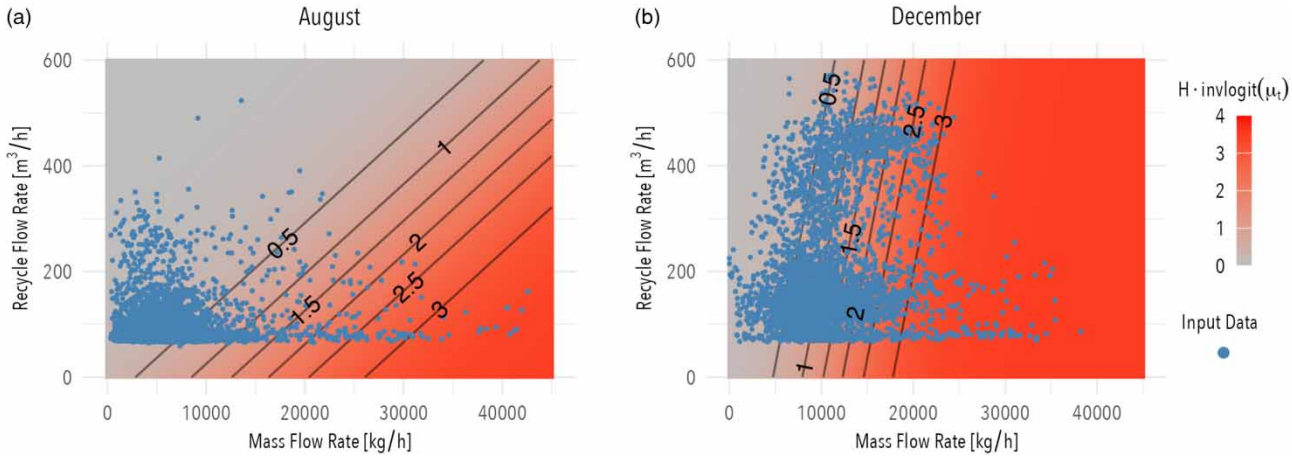

Figure 3 | The two images show contour plots for the estimated stationary sludge blanket $H_{in_{vlogit}}(\mu_t)$ based on the estimated parameters in Table 2 with input data shown as blue points.

expected the former behavior based on findings in the literature (Jones & Schuler 2010), although the authors mention that the effect is not observed at all treatment plants. The resulting form of the stationary sludge blanket $H_{invlogit}(\mu_t)$ function is plotted in Figures 3 and 10 as contours. A similar conclusion to what was just found with the scaled estimates may be derived from these plots. The slope of the contour lines on Kolding Central in December is much higher than in August, implying that along the y direction, which corresponds to an increasing recycle flow rate, the stationary blanket height varies little. Direct interpretation of the contour suggests that increasing the recycle flow rate from $100\,\mathrm{m}^3/\mathrm{h}$ to $400\,\mathrm{m}^3/\mathrm{h}$ decreases the sludge blanket by approximately 0.75 meters in December, but 2 meters in August. A similar exercise at Damhusåen yields a decrease in the sludge blanket by approximately 3 meters in December, and an identical 2 meters in August. Even though this latter proposed strength of the recycle flow is not based on an extrapolation of the input data as many input data are available in the y direction of Figure 10(b), it is regarded with a certain degree of suspicion. As already mentioned this conclusion is contrasted by the findings on the well-performing Kolding Central plant. In particular, inspecting the contour in December shows that for mass flow rates above $\approx 25,000\,\mathrm{kg/h}$ the recycle flow rate is practically unable to bring the stationary blanket below 3 meters, hence lowering the blanket may almost only be achieved by decreasing the plant inflow rate.

The standardized one-step-ahead residuals from the state estimation are shown in the various plots in Figures 4 and 11. It is clear that the residuals are colored, hinting that the assumption of normality does not hold. The quantile–quantile plot in particular bears evidence that the distribution has heavy tails, and there is significant auto-correlation present, in particular in the first lag. The former suggests that the residual distribution is better described by, e.g., a t-distribution, while the latter suggests adding another state into the system as a way to implement a time-delay, equivalent to going from an AR(1) to AR(2) process in discrete time. Several attempts were made to formulate an augmented system with an additional Ornstein-Uhlenbeck process for μ_t, with an identical structure to (8), but this did not improve the auto-correlation, and expanding the model was thus dropped. A challenge in adding original states to the system is not to introduce inputs which are impractical to forecast, or altogether inaccessible. The immediate conclusion here is clear, namely that the system is not expanded to a degree where the residuals show a satisfactory degree of whiteness, and future work should address this issue.

3.2. Predictive performance with perfect input forecast

The sludge blanket height predictions are made by integrating forward in time the moment differential equations of (8), without performing state updates based on the observations. The moment differential equations used in the Extended Kalman filter are given by:

$$\frac{\mathrm{d}m_t}{\mathrm{d}t} = f(m_t, u_t, \theta) = \frac{1}{\alpha}(\mu_t - m_t) \tag{16a}$$

$$\frac{\mathrm{d}v_t}{\mathrm{d}t} = 2\frac{\partial f}{\partial h}(m_t, u_t, \theta)v_t + g(m_t, u_t, \theta)^2 = -\frac{2}{\alpha}v_t + \sigma^2, \tag{16b}$$

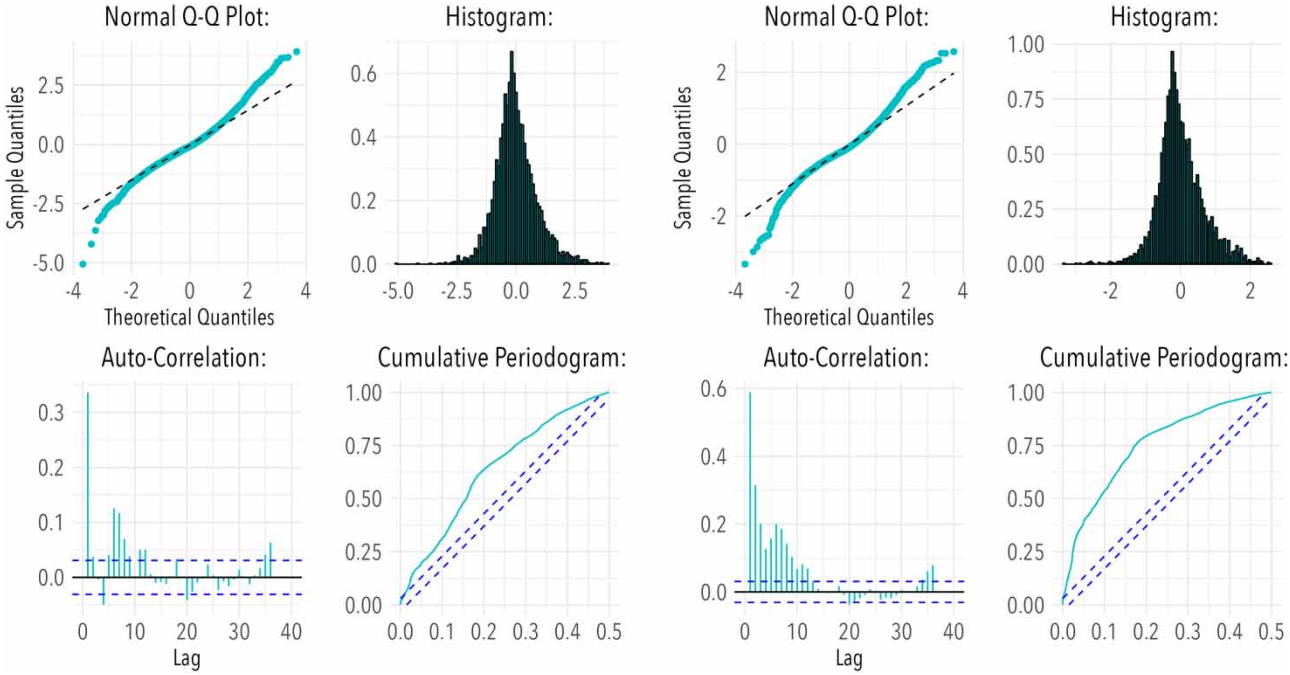

Figure 4 | Standard one-step-ahead residual analysis plots to verify the assumption of normality, for August and December at Kolding Central (left and right, respectively). The four plots shown are quantile–quantile plot (top left), histogram (top right), auto-correlation (bottom left) and cumulative periodogram (bottom right).

where $m(t)$ and $v(t)$ are the mean and variance, respectively, i.e. $m(t) = \mathbb{E}(h_t)$, $v(t) = \mathbb{V}(h_t)$. In the following we refer to a k-step prediction as the solution to (17) from time $t = t_i$ to t_{i+k} given some initial conditions $[m(t_i), v(t_i)] = [\hat{m}_i, \hat{v}_i]$, without updating the states based on observations. In the present case, due to the chosen data aggregation level each single step corresponds to 10 min i.e. a 1 h prediction is six steps. An example of model predictions are demonstrated in Figures 5 and 12 at Kolding Central and Damhusåen, respectively. The figures show 4-day (576 step) predictions of the sludge blanket height , where we assume known future inputs i.e. perfect forecasting of the input values. The number of observations that lie outside of the 95% prediction interval are shown in green (above the region) and red (below the region) on both plots. The 95% prediction intervals are recovered from (12) i.e. $\mathbb{V}(z_{t_k}) = \mathbb{V}(h_{t_k}) + \sigma_z^2$. The model parameters were estimated based on data from the past 30 days. In practical usage the typical forecast horizon of interest is between 2–10 h, so the 4-day time period chosen here is much beyond that, but was deliberately chosen to be exagerrated to showcase what we regard as great predictive capabilities, especially at Kolding Central. The prediction examples demonstrate a clear difference in performance between plants

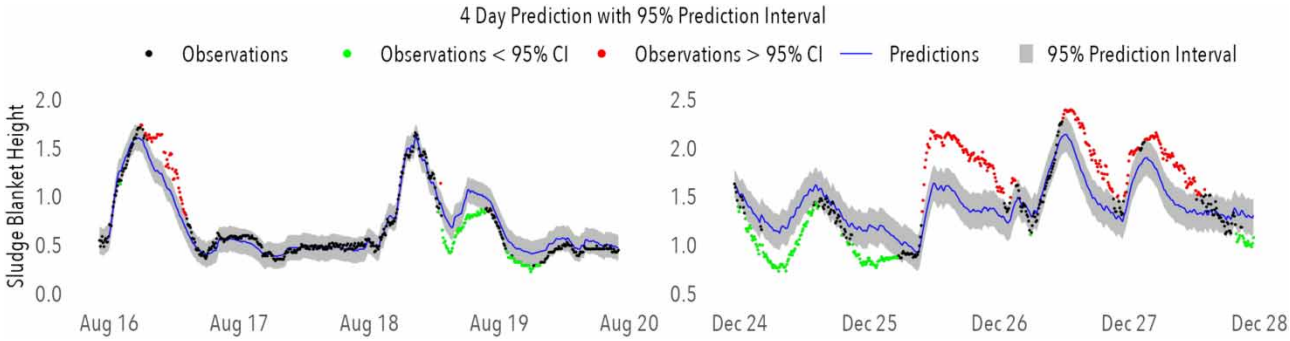

Figure 5 | A 4-day prediction of the sludge blanket height with associated 95\% prediction interval at Kolding Central for August (left) and December (right) for one selected clarifier. Also displayed are observations above and below the prediction interval in green and red, respectively.

and months. The performance at Kolding Central is relatively good for both months but clearly worse in December, with root-mean-square errors of 0.12 m and 0.31 m, respectively. The challenge in December is in particular seen to be the blanket height peak that occurs during December 25th, and generally the prediction uncertainty interval is too narrow. This is in correspondence with the fact that the proportion of outliers was found to be approximately 22% and 66%, respectively, thus the predictive distributions of the model appear to be too narrow, a challenge that we were unable to resolve (see Section 3.3), and which should be further addressed in future work. The performance at Damhusåen is worse with RMSEs of ≈ 0.38 m at both. The primary difference between the months is the broad uncertainty interval during December, caused by the much larger $\tilde{\sigma}_x$ there, which is reflected in the outlier percentages of 22% and 0%, respectively. The primary challenges during the two months are seen to be the peak on August 17th and the declining trend on December 6th.

The overall predictive performance analyses of the model are summarized in Figures 6 and 14 showing RMSEs for 60-step (10 h) predictions for all clarifiers based on periods with 'slow' (left plot) and 'fast' (right plot) dynamics, respectively. The predictions are made based on parameters estimated from the past 30 days of data. The 'fast' dynamics period is identical to that showcased in the 4-day prediction examples just presented, in Figures 5 and 12, and these are of primary interest since in general a persistence model should be good if the blankets are approximately constant. The prediction accuracy presented here should be regarded as an upper bound since known future input values are used, which corresponds to the unrealistic scenario with perfect forecasting of the mass inflow rate.

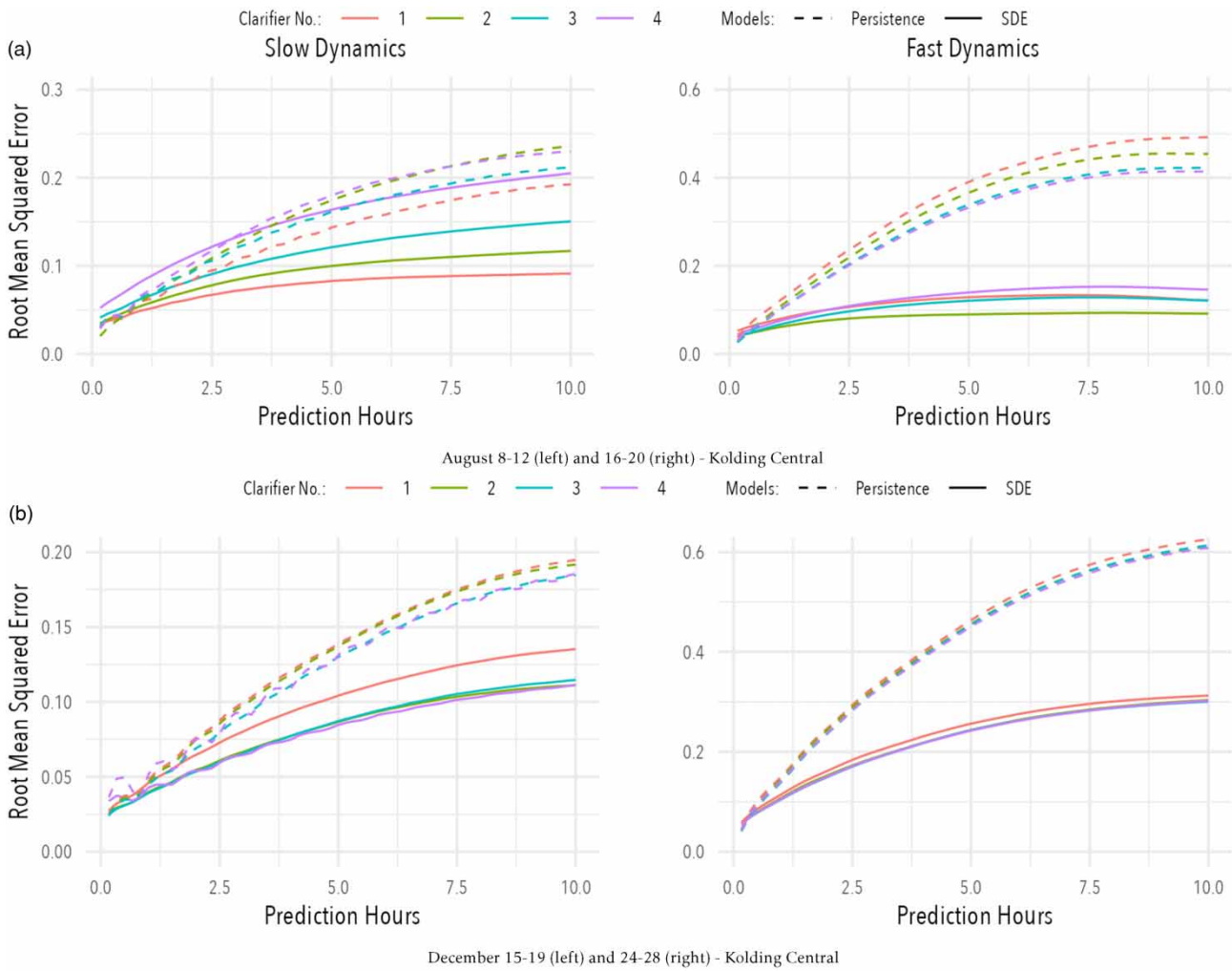

Figure 6 | The plots show root-mean-square errors for a 10 h prediction horizon, using known future inputs on selected 4-day periods where the blanket is either relatively stationary (left, "slow dynamics") or excited (right, "fast dynamics"). (a) August 8–12 (left) and 16–20 (right) – Kolding Central and (b) December 15–19 (left) and 24–28 (right) – Kolding Central.

The results overall show significant improvements in the RMSE over a persistence model for the 'fast dynamics' period with the exception of December on Damhusåen where results show little to no improvement (and even a worsening for Clarifier 1). The 'slow dynamics' period shows vastly different results at the two plants, with clear improvements at Kolding Central and much worse accuracy at Damhusåen. This latter result reveals that the model predictions do not agree with the stationary level of the observations as demonstrated in Figure 13 in Appendix B, and the model seems to wrongly predict minor blanket movement. The four missing curves in Figure 14(b) reveals challenges with missing data for this particular period, an issue that is indicative of the data conditions and variability of data between clarifiers at Damhusåen in general.

A comparison between the *average* RMSE's for the two models is made in Table 3 for three selected prediction hours for the 'fast dynamics time period'. We calculated the reported 'RRMSE' (Relative RMSE) and 'DRMSE' (Difference RMSE) as:

$$RRMSE_t = \frac{RMSE_t^{(\mathrm{SDE})} - RMSE_t^{(\mathrm{pers})}}{RMSE_t^{(\mathrm{pers})}} \cdot 100\% : \quad \mathrm{DRMSE} = RMSE_t^{(\mathrm{SDE})} - RMSE_t^{(\mathrm{pers})}, \tag{17}$$

where 'pers' is the persistence model and 'SDE' is our proposed model. The figures in the table further emphasizes the bad performance at Damhusåen in December, a strong and comparable performance at Damhusåen in August and Kolding Central in December, and a great performance at Kolding Central in August. We do note however that the DRMSE improvements are largest for Damhusåen in August although this may partially be explained by a particularly inaccurate persistence forecast.

We conclude the present section by offering an example of the effect of using non-perfectly forecasted inputs as a way to acknowledge that in practice the accuracy presented in the figures above will be reduced. It should be emphasized that the idea here is not to mimic reality exactly, but merely to demonstrate the predictions' sensitivity to the input signal. In the following we let $t = 0$ and $t = 1$ denote the start and end of the (10 h) prediction horizon. The authors have no knowledge of the existence of accurate suspended solids concentration forecasts so here we have assumed a persistence forecast i.e. $C_{f,t} = C_{f,0}$, but we note that a more sophisticated forecast should take into account the reductions that occur during high inflow rates. Focusing on the plant inflow forecast, we consider two different scenarios namely (1) a constant relative error, and (2) a linearly increasing relative error. Specifically, the former is constructed by scaling $Q_{f,t}$ with some constant K i.e. $Q_{f,t,K,\mathrm{constant}} = K \cdot Q_{f,t}$, while the latter uses a time-dependent scaling instead i.e. $Q_{f,t,K,\mathrm{linear}} = \bar{K}(t) \cdot Q_{f,t}$, where:

$$\bar{K}(t) = (1 - t) + tK. \tag{18}$$

This is a linear interpolation between $(t, K) = (0, 1)$ and $(t, K) = (1, K)$. That is, the relative error of the latter increases linearly in time, from no error at $t = 0$ to a relative error of K at $t = 1$, such that the error at $t = 1$ is identical to the error of the constant scaling scenario. It is clear from the comparison between these two signals, as shown in Figure 7, that the difference is quite large in the beginning, but miniscule towards the end of the horizon. The latter linear error scaling is intended to mimic a random walk whose variance grows linearly in time. The constant relative error forecast is clearly the more pessimistic of the two, and also the least natural, since we expect little initial error in our input forecasts. It should be noted that the chosen values for $K \in [0.5, 2]$ were used here because they produced appropriate uncertainties for a prediction horizon of 10 h, but for longer horizons those values should probably be altered. Using these two types of perturbed plant inflow signals, we calculate the RMSE for a range of scaling K for the two treatment plants during August and present the results in Figures 8 and 15.

Table 3 | Comparing the average root-mean-square errors of the persistence model with our proposed model for both treatment plants at selected prediction horizons

	Kolding Central				Damhusåen			
	RRMSE (%)		DRMSE		RRMSE (%)		DRMSE	
Pred. (hours)	August	December	August	December	August	December	August	December
2.5	−52	−39	−0.11	−0.11	−33	−3	−0.14	−0.03
5.0	−66	−46	−0.24	−0.21	−42	−14	−0.26	−0.08
10.0	−73	−50	−0.33	−0.31	−49	−16	−0.39	−0.12

A negative sign indicates that the proposed model has a lower RMSE and vice versa. The reader is referred to the bulk text for further details.

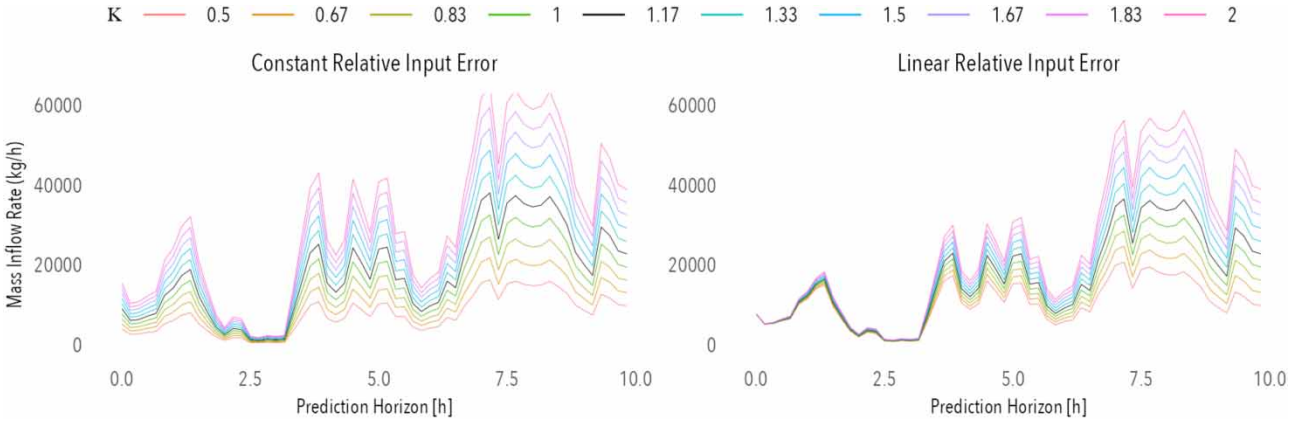

Figure 7 | A comparison between the constant relative error in the input signal (top), and one that increases linearly in time (bottom), for a 10-h horizon taken from August 18 00:00–10:00. The number K is the input scaling, and the black line with $K = 1$ is the observed input signal.

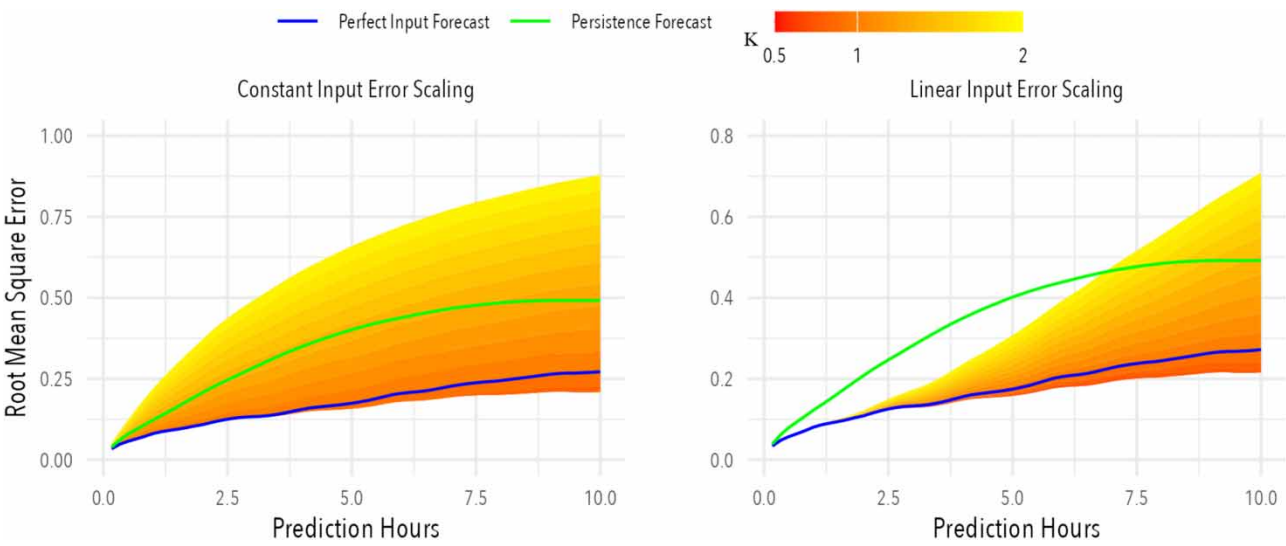

Figure 8 | The plots show the root-mean square error (RMSE) dependence on the input forecast error, for constant relative error (left) and linearly increasing relative error (right) for various scaling K on Kolding Central. Also shown is the RMSE for the perfect input forecast and for the persistence forecast.

Inspecting the figures we see, unsurprisingly that the errors in the forecast signal of the plant inflow has a large influence on the prediction accuracy. When using the linear input error scaling the model accuracy remains below that of the persistence model for the lowest and largest scaling for prediction horizons of approximately 0–10 h at Damhusåen and 0–7.5 h at Kolding Central. In contrast the accuracy is more challenged when using the constant input scaling where the accuracy decreases below that of the persistence model at scaling of $K \geq 1.5$ and $K \geq 1.7$ for Kolding Central and Damhusåen, respectively. It should be noted that while Damhusåen's RMSE distribution appears less sensitive relative to the persistence forecast the accuracy at Kolding Central remains highest in absolute terms for all scaling.

3.3. Discussion

We shall start by addressing performance, and differences therein, at the two treatment plants. First, the results presented in the previous section reveal that the performance in general is better in the summer month, although Kolding Central also showed good performance in December. For Damhusåen the prediction accuracy was significantly reduced in December with two out of six clarifiers performing worse than, or identical to, the persistence model, and the remaining four clarifiers

showing only marginal accuracy improvements. A possible explanation for this seasonal variation may be due to more complicated settling dynamics during winter which our proposed model is unable to capture well. It should be mentioned however that an additional challenge for Damhusåen in December was a (relatively) large number of missing observations. Secondly, it is evident when comparing the two plants that model performance at Kolding Central is superior. This fact is particularly emphasized by the accuracy for the 'slow dynamics' periods, which reveals that the model is able to capture the smaller blanket dynamics and the stationary blanket level very well. The finding that performance at Damhusåen is worse is not unexpected based on the large blanket variability at Damhusåen in the data presented in Figures 2 and 9. The model would only hypothetically be able to account for these individual sludge blanket differences if they were caused by differences in the recycle flow rate alone. The differences cannot arise from the mass inflow rate, since this signal is comprised of a plant-wide inflow (same for all 24 clarifiers at the plant) and a line-wide suspended solids concentration (same for all six clarifiers in a line). It is important to note that this challenge cannot be solved by measuring the two inputs on clarifier level since these unobtainable signals would then have to be forecasted. The modeling task at Damhusåen is further complicated by the fact that the distribution of the plant inflow to the operation lines (and clarifiers) is known to be non-uniform and time varying, which inevitably obscures parts of the correlation between inputs and observations. This flow distribution property is also what we believe to be the root cause of the large sludge blanket dispersion visible in the data for Damhusåen. Thirdly, the model stability was seen to be excellent at Kolding Central and very challenged at Damhusåen. In further details, we saw no period at Kolding Central were the model did not converge to a 'true' minimum, by which we mean an estimate whose maximum gradient component is sufficiently small. In contrast convergence to non-true optima or errors in the gradient/hessian calculations during optimization (which leads the optimization to a halt) were common when working with the Damhusåen data. This is inevitably tied to the data properties with larger dispersion, more frequent number of longer periods with missing observations, more missing observations in general, and the presence of sudden jumps in the blanket data. The stability was however drastically improved by increasing the observation variance (from $\sigma_z^2 = 0.05^2$ to $\sigma_z^2 = 0.10^2$), but errors and convergence to false minima remained for the more challenging data (such as December 27th–31st) that had many missing observations and larger jumps. In conclusion we have seen that high quality data without too many missing values and jumps and a well-controlled clarifier flow distributions is necessary in order for the model to be both stable and well-performing, especially during winter where more complicated settling dynamics may be present.

The present work lacks a satisfying treatment of the presented model's prediction uncertainty, although we briefly mentioned that the predictive distributions appear both underdispersed and overdispersed based on the results in Figures 5 and 12, respectively. We note that the outliers in the former figure are found primarily during periods where the blanket moves so it would seem sensible to have a diffusion function in (8) that incorporates this addition uncertainty. Various attempts were made to this end by expanding σ as a function of the inputs, e.g., $\sigma_t = \sigma Q_{f,t}^p$, $p \in [0, 1]$, with the interpretation that increased flow rates creates additional turbulence in the water column, but all of these attempts led to a decrease in the prediction accuracy, and were thus abandoned. A correct uncertainty quantification is very important both because this is an essential part of the underlying Kalman filter and heavily affects the individual likelihood contributions, but also for the purpose of predictive control where a natural control objective is to have a particular quantile remain under the top of the water column.

In regard to model training, a subtle challenge is the implicit emphasis on having higher prediction accuracy in the upper layers of the clarifier, where there is a greater risk of sludge overflow. The actuality of this challenge rests on the (unknown) assumption that the dynamics at higher levels are significantly different from those at lower levels. If that is the case then it may be difficult to learn the upper-layer dynamics in periods where the blanket is seldomly excited to those levels, and that may decrease prediction accuracy when it is needed most. This challenge leads us to ask the following two questions:

(1) How can we correctly measure the model prediction accuracy if the accuracy at lower levels is of little interest?
(2) What can be done in terms of model training to direct focus on performance in the upper clarifier layers?

We have addressed the former question by considering shorter prediction periods with substantial dynamics, as is the case in Figure 6 (top, right). Admittedly this is difficult, and even in the (carefully) selected period from August 12–16 the blanket is constant for a substantial period of time. In spite of this, estimation on peak-only blanket data (which we have omitted from the article) showed similar prediction accuracy, so we conclude that periods like August 12–16 are adequate representations of high blanket dynamics. For online estimation using a moving horizon of 2–3 weeks of data robust accuracy may be secured by including additional shorter periods with peak blanket events to retain knowledge of peak dynamics. It will however be

non-trivial to determine when such hand-picked periods are 'outdated' i.e. do not represent the current dynamics well due to changes in the clarifier conditions. With regards to the second question we propose to modify the observation variance function such that it decreases with increasing sludge blanket height. The log-likelihood due to the Kalman filter corresponds to a weighted least squares estimation so this suggestion amounts to increasing the weights on residuals related to higher blanket observations. This is similar to the procedure in Bergsteinsson *et al*. (2022) where a sigmoid function was used, although there the weighting was controlled by an input rather than a state. This suggestion does however introduce two hyperparameters describing the shape and location of the sigmoid function, and further complicates evaluation of the predictive accuracy since the model is expected to be accurate only in the upper layers.

A technical detail worthy of mentioning, although common in practice, is the fact that the objective function to be minimized is based on the likelihood of one-step-ahead residuals as opposed to using multi-steps-ahead. The latter would be more natural since it matches the intended purpose of the model which is to produce forecasts. In the present case the relevant steps-ahead would be 30–60 which corresponds to 5–10 h (10 min per step). The use of such an objective function will however lead to correlation between all residuals in the time series significantly complicating the likelihood computation. That being said, naive implementations that do not take residual correlation into account could be further investigated. Such an implementation is arguably no worse than the colored residuals we have presented in Figures 4 and 11.

These non-normal residual structures remain an issue which should be addressed. We have considered other observation noise distributions through state-transformations, in particular log and logit domains. A detailed account of these can be found in Appendix C. The conclusion of this work was, however, that there was only a slight reduction in the auto-correlation, but a large decrease in model stability. As a final comment we remark that since the model purpose is prediction accuracy improvements in the one-step-ahead residuals alone are not enough but must also be accompanied by improvements in accuracy since this is the ultimate objective.

Finally, for a future outlook on the perspectives here, we mention that Thilker *et al*. (2021) showed the possibility of incorporating input forecasts into a stochastic differential equation through additional states. In the present case it would be obvious to try to incorporate the plant inflow $Q_{f,t}$ into the model, although this will require additional observations, such as radar precipitation data. If successful, such an approach could generalize the modeling procedure across many wastewater treatment plants allowing a much easier implementation of the model and an associated clarifier predictive control strategy, limiting the required number of sensors and measurement devices.

4. CONCLUSION

A novel data-driven stochastic state space system for modeling and forecasting the sludge blanket height in secondary clarifiers have been presented. The model is trained on measurements of the sludge blanket height, and uses as inputs: (1) the mass inflow rate and (2) the clarifier recycle flow rate. The model prediction accuracy was observed to be great, relative to that of a persistence forecast, at a treatment plant with high data quality and well behaved clarifiers. The accuracy was more challenged at a treatment plant with lower data quality and uneven flow-distribution, in particular during winter, but the summer performance remained good. The model stability was similarly challenged at this more difficult-to-model plant, but was excellent on the former plant. The model residual analysis (one-step in-sample prediction errors) revealed a systematic pattern in the auto-correlation and deviated from the assumed normal distribution which indicates that there are missing drivers to be accounted for, but this issue could not be resolved. In conclusion, the presented model has the potential to be used in a model predictive control setting to improve secondary clarifier performance. As a concrete example a bypass algorithm could be developed whose objective is to maintain some prediction quantile of the sludge blanket height under the clarifier height, by allowing some fraction of the inflow to be bypassed. A similar strategy controlling the recycle flow rate, and trying to limit the effluent suspended solids could be used. Consequently the authors regard the presented work as a step towards the development of tomorrow's modern data-driven wastewater treatment plants.

ACKNOWLEDGEMENTS

The data analysed in this paper were provided by Biofos (Damhusåen) and BlueKolding (Kolding Central). The work was partly funded by Biofos A/S and Krüger A/S in relation to the first author's Ph.D studies. The work on developments of the used gray box modeling tools is supported by *ARV* (EU H2020 101036723), and *ELEXIA* (EU Horizon Europe

101075656). The authors acknowledge a conflict of interest insofar as the third and fourth co-authors are employed at Krüger A/S.

CONFLICT OF INTEREST

The authors have a conflict to declare.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

REFERENCES

Anderson, N. E. & Gould, R. H. 1945 Design of final settling tanks for activated sludge (with discussion). *Sewage Works Journal* **17** (1), 50–65.

Balslev, P., Nickelsen, C. & Lynggard-Jensen, A. 1994 On-line flux-theory based control of secondary clarifiers. *Water Science Technology* **30** (2), 209–218.

Bergsteinsson, H. G., Vetter, P. B., Møller, J. K. & Madsen, H. 2022 Estimating temperatures in a district heating network using smart meter data. *Energy Conversion and Management* **269**, 116113.

Brok, N. B., Madsen, H. & Jørgensen, J. B 2018 Nonlinear model predictive control for stochastic differential equation systems. *IFAC-PapersOnLine* **51**, 430–435.

Bürger, R., Diehl, S. & Nopens, I. 2011 A consistent modelling methodology for secondary settling tanks in wastewater treatment. *Water Research* **45**, 2247–2260. doi: 10.1016/j.compchemeng.2012.02.016.

Bürger, R., Diehl, S., Faras, S. & Nopens, I. 2012 On reliable and unreliable numerical methods for the simulation of secondary settling tanks in wastewater treatment. *Computers and Chemical Engineering* **41**, 93–105.

Clarcq, J. D., Devisscher, M., Boonen, I., Vanrolleghem, P. A. & Defrancq, J. 2003 A new one-dimensional clarifier model-verification using full-scale experimental data. *Water Science and Technology* **47** (12), 105–112.

Coe, H. & Clevenger, H. 1916 Methods for determining the capacities of slime settling tanks. *Transactions of the American Institute of Mining, Metallurgical and Petroleum Engineers* **50**, 356–384.

Ekama, G., Barnard, J., Günthert, F., Krebs, P., McCorquodale, J., Parker, D. & Wahlberg, E. 1997 *Secondary Settling Tanks. Theory, Modelling, Design and Operation*. International Associated of Water Quality.

Fitch, B. 1966 Current theory and thickener design. *Industrial and Engineering Chemistry* **58** (10), 18–28.

Gay, D. M. 1990 *Usage Summary for Selected Optimization Routines*. Tech. Rep. AT&T Bell Laboratories.

Guyonvarch, E., Ramin, E., Kulahci, M. & Plósz, B. G. 2015 icfd: Interpreted computational fluid dynamics - degeneration of cfd to one-dimensional advection-dispersion models using statistical experimental design - the secondary clarifier. *Water Research* **83**, 396–411.

Guyonvarch, E., Ramin, E., Kulahci, M. & Plósz, B. G. 2020 Quantifying the sources of uncertainty when calculating the limiting flux in secondary settling tanks using icfd. *Water Science and Technology* **81** (2), 241–252. doi: 10.2166/wst.2020.090.

Hach 2023 *Hach Sludge Blanket Sensor*. Available at: https://www.hach.com/p-sonatax-sc-sludge-level-and-sludge-height-probe-with-wiper-stainless-steel/LXV431.99.00002#specifications (accessed 21 July 2023).

Hamilton, J., Jain, R., Antoniou, P., Svoronos, S. & Koopman, B. 1992 Modeling and pilot-scale experimental verification for predenitrification process. *Journal of Environmental Engineering* **118** (1), 551.

Iacus, S. M. 2008 *Simulation and Inference for Stochastic Differential Equations: With R Examples*. Springer, New York, USA. doi: 10.1007/978-0-387-75839-8.

Jazwinski, A. 1970 *Stochastic Processes and Filtering Theory*. Dover Publications, New York, USA.

Jeppsson, U. & Diehl, S. 1996 An evaluation of a dynamic model of the secondary clarifier. *Water Science Technology* **34**, 19–26.

Jones, P. A. & Schuler, A. J. 2010 Seasonal variability of biomass density and activated sludge settleability in full-scale wastewater treatment systems. *Chemical Engineering Journal* **164**, 16–22. doi: 10.1017/S0027763000012216.

Junker, R. G., Kallesoe, C. S., Real, J. P., Howard, B., Lopes, R. A. & Madsen, H. 2020 Stochastic nonlinear modelling and application of price-based energy flexibility. *Applied Energy* **275**, 115096. doi: 10.1016/j.apenergy.2020.115096.

Keinath, T. M. 1985 Operational dynamics and control of secondary clarifiers. *Water Pollution Control Federation* **57** (7), 770–776.

Keinath, T. M. 1990 Diagram for designing and operating secondary clarifiers according to the thickening criterion. *Water Pollution Control Federation* **62** (3), 254–258.

Keinath, T. M., Asce, M., Ryckman, M. D., Dana, C. H. & Hofer, D. A. 1977 Activated sludge-unified system design and operation. *Journal of The Environmental Engineering Division* **103** (5), 829–849.

Kos, P. 1977 Thickening of water-treatment-plant sludges. *American Water Works Association* **69** (5), 272–282.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. & Bell, B. M. 2016 Tmb: Automatic differentiation and laplace approximation. *Journal of Statistical Software* **70** (5), 1–21. doi: 10.18637/jss.v070.i05.

Kynch, G. J. 1952 *A Theory of Sedimentation*. Department of Mathematical Physics, Birmingham University.

Madsen, H., Juhl, R. & Møller, J. K. 2015 *Continuous Time Stochastic Modeling in R – User's Guide and Reference Manual*.

Møller, J. K. 2011 *Stochastic State Space Modelling of Nonlinear systems - With application to Marine Ecosystems*. English. Doctoral Dissertation.

Møller, J. K., Bergmann, K. R., Christiansen, L. E. & Madsen, H. 2012 Development of a restricted state space stochastic differential equation model for bacterial growth in rich media. *Journal of Theoretical Biology* **305**, 78–87.

Plosz, B. G., De Clercq, J., Nopens, I., Benedetti, L. & Vanrolleghem, P. A. 2011 Shall we upgrade one-dimensional secondary settler models used in WWTP simulators? – An assessment of model structure uncertainty and its propagation. *Water Science and Technology* **63** (8), 1726–1738. ISSN: 0273-1223, doi: 10.2166/wst.2011.412.

Qiu, Y., Hug, T., Wágner, D. S., Smets, B. F., Valverde-Pérez, B. & Plósz, B. G. 2023 Dynamic calibration of a new secondary settler model using cand. microthrix as a predictor of settling velocity. *Water Research* **246**, 120664. ISSN: 0043-1354. doi: 10.1016/j.watres.2023.120664.

R Core Team 2023 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

R-Stats-Documentation 2023 nlminb. Available from: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/nlminb.html (accessed 21 July 2023).

Ramin, E. 2014 *Modelling of Secondary Sedimentation Under Wet-Weather and Filamentous Bulking Conditions*. English. Doctoral dissertation.

Stentoft, P. A., Munk-Nielsen, T., Vezzaro, L., Madsen, H., Mikkelsen, P. S. & Møller, J. K. 2019 Towards model predictive control: Online predictions of ammonium and nitrate removal by using a stochastic ASM. *Water Science and Technology* **79** (1), 51–62.

Stentoft, P. A., Munk-Nielsen, T., Møller, J., Madsen, H., Pérez, V., Mikkelsen, P. & Vezzaro, L. 2021 Prioritize effluent quality, operational costs or global warming?–Using predictive control of wastewater aeration for flexible management of objectives in WRRFs. *Water Science and Technology* **196**, 116960.

Stukenberg, J. R., Rodman, L. C. & Touslee, J. E. 1983 Activated sludge clarifier design improvements. *Water Pollution Control Federation* **55** (4), 341–348.

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

Tekippe, R. J. & Bender, J. H. 1987 Activated sludge clarifiers: Design requirements and research priorities. *Water Pollution Control Federation* **59** (10), 865–870.

Thilker, C. A., Madsen, H. & Jørgensen, J. B. 2021 Advanced forecasting and disturbance modelling for model predictive control of smart energy systems. *Applied Energy* **292**, 116889. doi: 10.1016/j.apenergy.2021.116889.

Thompson, D., Chapman, D. T. & Murphy, K. L. 1989 Step feed control to minimize solids loss during storm flows. *Research Journal of the Water Pollution Control Federation* **61** (11), 1658–1665.

Thygesen, U. H. 2023 *Stochastic Differential Equations for Science and Engineering*. Taylor and Francis, Boca Raton, FL, USA. doi: 10.1201/978-1-003-27756-9.

Vesilind, P. A. 1974 *Treatment and Disposal of Wastewater Sludges*. Ann Arbor Science Publ, Michigan, USA.

Vetter, P. B. 2020 *Modelling and Control of Sludge Blanket using Stochastic Differential Equations*. Master's Thesis, Technical University of Denmark.

Watts, R. W., Svoronos, S. A. & Koopman, B. 1996 One-dimensional modeling of secondary clarifiers using a concentration and feed velocity-dependent dispersion coefficient. *Water Research* **30** (9), 2112–2124.

Wickham, H. 2016 *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York. ISBN: 978-3-319-24277-4. https://ggplot2.tidyverse.org.

Xie, Y. 2014 knitr: a comprehensive tool for reproducible research in R. In: *Implementing Reproducible Computational Research* (V. Stodden, F. Leisch, & R. D. Peng, eds.). Chapman and Hall/CRC. ISBN 978–1466561595.

Xu, Q., Luo, X., Xu, C., Wan, Y., Xiong, G., Chen, H., Zhou, Q., Yan, D., Li, X., Li, Y. & Liu, H. 2022 The whole process cfd numerical simulation of flow field and suspended solids distribution in a full-scale high-rate clarifier. *Sustainability* **14** (17), 10624. doi: 10.3390/su141710624.

Zhou, P. & Li, Z. 2023 Arbitrary polynomial chaos expansion for uncertainty analysis of the one-dimensional hindered-compression continuous settling model. *Journal of Water Process Engineering* **52**, 103489. ISSN: 2214-7144. doi: 10.1016/j.jwpe.2023.103489.

Zhu, H. 2021 *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4. https://CRAN.Rproject.org/package=kableExtra.

First received 22 December 2023; accepted in revised form 30 June 2024. Available online 13 July 2024