Using Bayesian Inference and Flowpipe Construction to Bound Predictions of Biogas Production at Wastewater Treatment Plants

Fletcher Chapin*, Ankur Varma†, Samuel Akinwande*,
Meagan Mauter* and Sriram Sankaranarayanan†.

(*) Stanford University, Stanford, USA
(†) University of Colorado Boulder, USA.

Abstract. In this paper, we use a novel combination of probabilistic programming and flowpipe construction to predict bounds on future biogas production for a wastewater treatment plant given operational data from the past. The operation of the anaerobic digester of a wastewater treatment plant is modeled through an Ordinary Differential Equation (ODE) model with unknown parameters and unobservable internal states. We are given data from the plant's operation that includes the daily measurement of the incoming waste volumes and concentrations along with the volume of biogas produced. We formalize our problem as first estimating the unknown parameters and initial conditions using Bayesian inference, such that the past behavior of the system is "compatible" with the observed data. Next, we propagate those input parameter estimates forward using flowpipe construction. To enable rapid and accurate flowpipe construction, we exploit the monotonicity property of the dynamical model of the plant. The procedure yields an over-approximation of the upper and lower bounds on biogas production, given the inputs. As a result, it can be used to formally bound future predictions that might inform facility operations. We implemented this procedure using a first-order kinetics model of hydrolysis to model the anaerobic digester of a real-world case study facility. We demonstrate how this method constructs realistic bounds for biogas prediction from the historical data that contain the actual ground-truth data 100% of the time. Our approach outperforms the standard approach that computes a posterior predictive distribution from samples both in terms of time and accuracy.

Keywords: Cyber-Physical Systems, Reachability Analysis, Flowpipe Construction, Monotone Systems, Waste-Water Treatment and Bayesian Inference.

1 Introduction

This work uses a combination of probabilistic programming and reachability analysis to predict probabilistic bounds on the future biogas production of wastewater treatment plants (WWTP). WWTPs involve a *digester* that employs a process of anaerobic digestion to convert various input waste streams into biogas that can be used as a source of renewable energy. The operation of WWTPs is rife with many sources of uncertainties. The process of digestion can be captured (approximately) by ordinary differential equation (ODE) models with uncertain rate parameters, and unobservable state variables. There is also uncertainty in the data, including the amount and concentration of wastes in the various input streams. Finally, data at WWTPs is measured at varying frequencies and accuracies. Some data is measured using real-time sensors, while other data is measured every few days by a laboratory technician. The differing frequency and accuracy in these measurements adds additional uncertainty to an already underspecified problem. Given all of these uncertainties, being able to obtain formal bounds on predictions of future conditions, such as anaerobic digester biogas production, would be useful to WWTP operators. For instance, the operator may wish to predict the biogas production over the near term for various future input scenarios.

In this paper, we use ideas from formal verification of cyber-physical systems (CPS) and probabilistic programming to analyze the operation of WWTPs. We wish to construct upper and lower bounds for the future biogas production of an anaerobic digester at a WWTP such that for an input probability level γ, the future biogas production should lie within the computed bounds with probability at least γ. Our method is a two-step approach. First, we use Bayesian modeling and inference to estimate a range of uncertain input parameters whose posterior probability exceeds γ [38]. Bayesian modeling and inference consists of specifying a generative model of the process of biogas production from the input waste streams. *Probabilistic programming* languages such as `PyRO` [10], `Stan` [13], `Anglican` [48] and `Turing.jl` [25] are commonly used to specify such models. We refer the reader to a survey by De Meent et al [39] or the monograph edited by Barthe et al [8] for a detailed introduction to probabilistic programming concepts. Our work uses the Julia programming language based `Turing.jl` to specify the model and perform inference [25]. The process of inference yields samples from the posterior distribution over the unknown parameters and initial conditions, conditioned on the observed data from plant operation. Second, we extract the *credible intervals* (the Bayesian analog of confidence intervals) over the input parameters and initial states such that the (sample-estimated) posterior probability of these intervals is at least γ [32]. Finally, given a planned future set of inputs, we perform reachability analysis over the differential equation model. Given a set of initial states and unknown parameters, reachability analysis constructs a flowpipe that is an over-approximation of all states reached over some finite time horizon of interest [4, 19]. By exploiting the monotonicity properties of the WWTP model, we show how to perform reachability analysis. Our approach here is a simplified version of a framework for reachability analysis proposed by Meyer et al [40]. This approach reduces reachability analysis to a simulation of the upper and lower bound trajectories of the system, showing that all behaviors are contained “in between” these trajectories. As a result, we obtain an enormous speedup over existing approaches.

We evaluate our two-step approach against a standard approach based on simulation of the samples from the posterior over a dataset that spans three months of operational data consisting of waste inputs and gas production outputs for a WWTP in the San Francisco Bay Area. We evaluate our approach by splitting the data into 8-day segments wherein parameters and initial conditions are obtained by running Bayesian inference over data from 3 days and then using data from the subsequent 5 days as a prediction period. We assess the flowpipes constructed for various credible intervals against the ground truth data. In addition to measuring prediction accuracy, we also compare the flowpipe bounds and computational times against those obtained through the posterior simulation approach. We show that our approach is more accurate in terms of containing ground truth data within the predicted bounds but also quite fast when compared to posterior simulations. However, we also note that our approach to deriving credible intervals can be quite conservative, leaving room for future improvements. The bounds predictions obtained using our approach could be used by plant operators to formulate a cost-minimizing energy management plan, especially at facilities with highly variable electricity tariffs [15], gas storage resources like a biogas storage tank [11, 16], or energy storage resources like a Li-ion battery [41, 11, 16].

1.1 Related Work

Significance and challenges: Recent research has highlighted the carbon emission reductions [43, 16] and electricity bill savings [41, 50, 11] potential of flexible operation of wastewater treatment plants (WWTPs). However, operational challenges make it difficult to realize those benefits in the real world [47]. Since WWTPs are critical infrastructure, facility operators are understandably risk-averse to modifying operations. Formal guarantees around predictions could help build operator confidence around novel control approaches such as model predictive control.

Reachability Analysis: This work builds upon the existing literature on computing reachable sets for cyber-physical systems described by ordinary differential equations [4, 19]. Reachability analysis techniques estimate bounds on the solutions to a differential equation model over a finite time horizon, given the set of initial conditions and unknown parameters. The problem of reachability analysis is known to be undecidable even for linear systems. However, recent approaches have provided computationally efficient and precise bounds to prove properties for systems with billions of state variables [6]. Unfortunately, these approaches are not directly applicable to our case due to the presence of uncertain parameters. Approaches that can handle non-linear systems can be used for our application. For non-linear systems, we have numerous approaches including Taylor-model based verified integration [9, 17, 18], or polynomial zonotopes [2], to mention a few. However, these approaches can be computationally expensive. Our work here uses a specialized approach that exploits monotonicity properties of the underlying model [40]. As a result, we obtain a precise bound that is also computationally inexpensive. Prior work has addressed the problem of reachability analysis for models with uncertainty by employing sampling-based approaches to reachability analysis as in [26, 36].

Sampling and Inference in Verification: The use of sampling and hypothesis testing to prove properties of stochastic systems has been explored extensively in the past through the statistical model checking approach [33, 22]. This has been applied in many domains, including biological systems [51]. Our approach is sample-based since we employ a Bayesian inference procedure that uses Monte Carlo methods to draw samples from the posterior [39]. Although inference approaches such as Hakaru can characterize the posterior distribution precisely [42], they are restricted to a relatively small class of programs, and cannot handle those involving differential equation models. As a result, our work employs Monte Carlo approaches with asymptotic guarantees that assume a sufficiently large number of samples. We emphasize that such an approach is very common in the probabilistic programming literature, wherein the properties of the posterior are inferred from a "large enough" number of samples. Our approach is in direct contrast to the standard statistical approach of constructing "posterior predictive samples" by simulating the posterior samples. However, we observe in our empirical evaluation that such an approach is relatively expensive since it involves simulation of a large number of samples and at the same time loses accuracy since it is unable to generalize effectively from the samples.

Prior work has explored the combination of reachability analysis and statistical inference for predictive monitoring of a variety of systems including road vehicles [3], stochastic processes [12], and human models [7]. Of these, our work is most closely related to the work of Chou et al [20] in which the authors apply Bayesian approximation methods to estimate forward projection sets for a vehicle model.

2 Problem Statement and Approach

Ordinary differential equations (ODEs) are used to model a variety of natural and engineered systems in domains ranging from physics to ecology. Let $\mathbf{x} \in \mathbb{R}^n$ represent a vector of state variables $\mathbf{x} = (x_1, \ldots, x_n)$ and $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ be a Lipschitz continuous function wherein $f(\mathbf{x}, \theta)$ is dependent on parameters $\theta \in \Theta \subseteq \mathbb{R}^m$. An ODE is of the form $\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \theta)$, wherein the function f (also known as the vector field) maps each state $\mathbf{x}$ and parameter $\theta \in \Theta$ to a derivative $f(\mathbf{x}, \theta)$. The solution to the ODE given fixed parameter values $\theta \in \Theta$ and initial conditions $\mathbf{x}(0)$ is a trajectory $\varphi : [0, T) \to \mathbb{R}^n$ for some time horizon $T > 0$ such that (a) $\varphi(0) = \mathbf{x}(0)$ and (b) for all $t \in [0, T)$, $\frac{d\varphi}{dt} = f(\varphi(t), \theta)$. In other words, the derivative of the solution satisfies the differential equation for a fixed θ. Since f is assumed to be Lipschitz continuous, we know that the solution exists and is unique. In this paper, we will work with ODEs with inputs. Let $\mathbf{u} \in \mathbb{R}^k$ represent a vector of k inputs.

An ODE with inputs drawn from a domain $U \subseteq \mathbb{R}^k$ has the form $\frac{d\mathbf{x}}{dt} = f(\mathbf{x}, \mathbf{u}, \theta)$. In many ODE models of physical systems (including that studied here), the internal state $\mathbf{x}$ is not directly measurable. We will assume that $\mathbf{y} = g(\mathbf{x})$ is a measurable output. Given a fixed parameter $\theta \in \Theta$ and an input signal $\xi : [0, T] \to U$, the corresponding trajectory φ for initial condition $\mathbf{x}(0)$ satisfies $\varphi(0) = \mathbf{x}(0)$ and $\frac{d\varphi}{dt} = f(\varphi(t), \xi(t), \theta)$.

For notational convenience, given a set of values $\mathbf{u}_1, \ldots, \mathbf{u}_t \in \mathbb{R}^k$, we write $\mathbf{u}(1, \ldots, t)$ to denote the function $[0, t) \to \mathbb{R}^n$ wherein $\mathbf{u}(\tau) = \mathbf{u}_i$ if $i - 1 \le \tau < i$.

Definition 1 (Problem Statement). *The problem of predicting future range of possible biogas production given past data and the model structure is as follows:*

Inputs: *Model structure* $\langle f, \Theta, U, g\rangle$, *past data* $\langle \mathbf{u}(1, \ldots, t), \mathbf{y}(1, \ldots, t)\rangle$, *future planned input signal* $\mathbf{u}(t + 1, \ldots, t + k)$, *and confidence level* $\gamma \in (0, 1)$.

Outputs: *Credible interval for parameters :* $[\theta_{lo}, \theta_{hi}]$ *and future output bounds* $[\mathbf{y}_{lo}(\tau), \mathbf{y}_{hi}(\tau)]$, *wherein* $\mathbf{y}_{lo}, \mathbf{y}_{hi} : [t, t+k) \to \mathbb{R}^{|\mathbf{y}|}$, *such that* $\forall \tau \in [t+1, t+k), \mathbf{y}_{lo}(\tau) \le \mathbf{y}_{hi}(\tau)$ *are bounds on the future outputs of the model.*

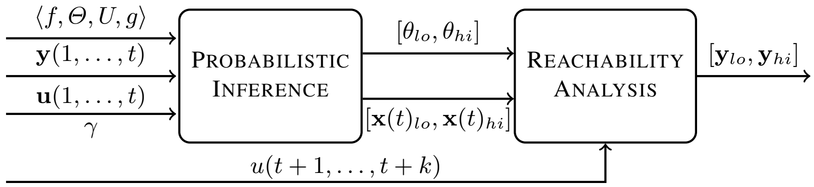


Fig. 1. Overall approach at a glance using a combination of probabilistic inference to estimate the parameter and initial condition posterior ranges while using reachability analysis to estimate bounds on future outputs.

The specific wastewater treatment model is detailed in section 3. It has $n = 7$ state variables, $k = 10$ inputs, $m = 7$ unknown parameters and a scalar ($|\mathbf{y}| = 1$) observable output which represents the gas production. Each parameter is known to belong to a large interval of possible values and the states of the plant $\mathbf{x}$ cannot be observed. As a result, the problem has two aspects to it: (a) given past data, find out possible values of parameters θ and possible initial states $\mathbf{x}(0)$; and (b) given future inputs, and the set of possible values of $\theta, \mathbf{x}(0)$, predict bounds on the outputs $\mathbf{y}$.

Figure 1 summarizes the overall approach to the problem, which consists of two parts: (a) we run Bayesian inference through a probabilistic programming framework to compute posterior credible intervals over the unknown parameters and final states of the model; and (b) we perform a reachability analysis over the unknown parameters and initial conditions given the future inputs to compute the overall bounds.

Bayesian Inference: Bayesian inference inputs (a) the generative model given by the probability distribution $\mathbb{P}(\mathbf{y}(1,\ldots,t) \mid \theta, \mathbf{x}(0), \mathbf{u}(1,\ldots,t))$ which represents the probability of observing the output data, given the inputs to the plant, the unknown parameters θ and initial state $\mathbf{x}(0)$ for the model; and (b) the *prior distribution* $\pi(\theta, \mathbf{x}(0))$ with support over the set $\Theta \times \mathbb{R}^n$. We seek to compute a representation of the posterior distribution $\mathbb{P}(\theta, \mathbf{x}(0) \mid \mathbf{y}(1,\ldots,t), \mathbf{u}(1,\ldots,t))$ using Bayes' rule:

$$\mathbb{P}\left(\theta, \mathbf{x}(0) \mid \mathbf{y}(1,\ldots,t), \mathbf{u}(1,\ldots,t)\right) \propto \mathbb{P}\left(\mathbf{y}(1,\ldots,t) \mid \theta, \mathbf{x}(0), \mathbf{u}(1,\ldots,t)\right) \pi(\mathbf{x}(0), \theta)\,.$$

There are many computational approaches to Bayesian inference, including techniques such as Markov chain Monte Carlo (MCMC), sequential Monte Carlo (SMC), Belief Propagation (BP), and Variational Inference (VI) [39, 8]. Tools such as `Turing.jl` [25], `PyRo` [10] and `Stan` [13] support Bayesian inference by specifying the model, and the prior distribution as programs in a domain specific language called a *probabilistic programming language*.

Our framework uses probabilistic programming to compute samples from the posterior distributions for $\mathbb{P}(\theta, \mathbf{x}(0) \mid \mathbf{y}(1,\ldots,t), \mathbf{u}(1,\ldots,t))$. Given the posterior and a confidence level $\gamma \in (0, 1)$, we extract *credible intervals* over $[\theta_{lo}, \theta_{hi}] \times [\mathbf{x}(t)_{lo}, \mathbf{x}(t)_{hi}]$ for $\theta, \mathbf{x}(t)$, such that the probability of drawing a sample from the credible interval is at least γ.

Reachability Analysis: We wish to predict bounds on the output $[\mathbf{y}_{lo}, \mathbf{y}_{hi}]$ such that for any choice of parameters $\theta \in [\theta_{lo}, \theta_{hi}]$, an initial state $\mathbf{x}(0) \in [\mathbf{x}(0)_{lo}, \mathbf{x}(0)_{hi}]$, an input signal $\mathbf{u}(1, \ldots, t+k)$, the resulting trajectories $\varphi : [0, t+k) \to \mathbb{R}^n$ is such that $\forall t \in [0, t+k), g(\varphi(t)) \in [\mathbf{y}_{lo}(t), \mathbf{y}_{hi}(t)]$. In other words, the reachability analysis returns a bound that accounts for all possible outputs that can be observed for the input parameter and initial condition ranges.

Overall Result: The Bayesian inference yields the following guarantee:

Theorem 1. *Assume that the observed data was generated by an instance of the model with parameter $\theta \in \Theta$ and $\mathbf{x}(0) \in \mathbb{R}^n$ that are sampled according to the prior distribution $(\theta^*, \mathbf{x}^*(0)) \sim \pi(\theta, \mathbf{x})$. Suppose, that we run a Bayesian inference procedure that generates N posterior samples. As $N \to \infty$, the probability that the $\theta^*, \mathbf{x}^*(0)$ lie within the credible intervals $[\theta_{lo}, \theta_{hi}]$ and $[\mathbf{x}(0)_{lo}, \mathbf{x}(0)_{hi}]$ is at least γ.*

The reachability analysis approach yields the following guarantee:

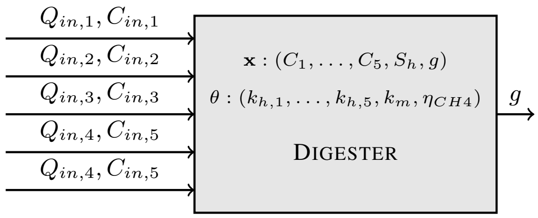
Theorem 2. *For any $\theta \in [\theta_{lo}, \theta_{hi}]$ and initial condition $\mathbf{x}(0) \in [\mathbf{x}(0)_{lo}, \mathbf{x}(0)_{hi}]$, the output obtained from the resulting trajectory $\varphi : [0, t+k) \to \mathbb{R}^n$ is contained in the computed reachable output bounds $[\mathbf{y}_{lo}, \mathbf{y}_{hi}]$:*

$$\forall\, t \in [0, t+k),\ \mathbf{y}_{lo}(t) \leq g(\varphi(t)) \leq \mathbf{y}_{hi}(t)\,.$$

Theorem 1 combined with the soundness guarantee of the reachability analysis from Theorem 2 yields the result that the bounds computed by our approach capture all possible future outputs of the system with probability at least γ.

3 Wastewater Treatment Model

The goal of our wastewater treatment model is to compute the biogas production in m^3/day based on the influent (incoming waste stream) concentrations and flow rates. We use a simplified first-order kinetics reaction to model the biogas production of an anaerobic digester. The anaerobic digester is modeled as a continuously stirred tank reactor (CSTR). Anaerobic digestion consists of three main steps. First, sludges and other wastes are hydrolyzed into simpler fats, acids, and proteins. Second, those simpler organic compounds are fermented into acetate and H_2. Third, acetate and H_2 are converted into CH_4 and CO_2 by acetotrophic or hydrogenotrophic methanogens [44]. Recent research has shown that hydrolysis can be rate limiting when there are high concentrations of volatile solids (VS) [44]. For example, Mahmood et al. found that hydrolysis was the rate-limiting step in the presence of high phosphine concentrations [37]. Since the rate-limiting step of anaerobic digestion depends on the wastewater's composition, our model incorporates two steps of the anaerobic digestion process: hydrolysis of complex wastes into simpler organic compounds and methane fermentation of those simple organics into CH_4. In other words, we combine fermentation and methanogenesis into a single step by assuming that methanogenesis is rate limiting. We model both steps using first-order kinetics models.



Inputs

1	Primary Slude
2	Waste Activated Sludge (WAS)
3	Food Waste
4	Fats, Oils and Greases
5	Scum

Output

g	Gas Production

Fig. 2. Diagram of anaerobic digester at a wastewater treatment plant. The variables associated with each flow are labeled.

Figure 2 shows the overall model of the digester in terms of a block diagram. The model has 7 state variables $C_1, \ldots, C_5$ representing the concentration of the various wastes coming from the input streams numbered $1, \ldots, 5$ whose meanings are explained in Figure 2, S_h representing the overall chemical oxygen on demand (COD), and the gas production g. The detailed explanation of the state variables, parameters and the differential equations is provided through the rest of this section.

3.1 Hydrolysis

We model reactions using first-order kinetics, meaning that the reaction rate is linearly dependent on the concentration of the reactant [49]. For our problem, there are j heterogeneous waste streams into the reactor. As a CSTR, we can assume that the effluent waste stream is fully mixed throughout the reactor. Let $Q_{in,j}$ be the influent flow rate of the j^{th} waste stream in m^3/day, Q_{out} be the effluent flow rate of the well mixed sludge in m^3/day, V be the total volume of wastewater in the reactor in m^3, $C_{in,j}$ be the influent volatile solids (VS) concentration of the j^{th} waste stream in mg / L, C_j be the VS concentration of the j^{th} constituent within the reactor in mg / L, and $k_{h,j}$ be the hydrolysis rate of the j^{th} waste stream in 1/day. Then, we can use mass balance to define dC_j/dt, the rate of change of the j^{th} waste concentration, in mg/L/day as

$$\frac{dC_j}{dt} = \frac{Q_{in,j}}{V} C_{in,j} - (\frac{Q_{out}}{V} + k_{h,j}) C_j \,, \tag{1}$$

wherein, the term $\frac{Q_{in,j}}{V} C_{in,j}$ represents the influent flow of waste stream j, $\frac{Q_{out}}{V} C_j$ represents the effluent flow, and $k_{h,j} C_j$ represents the mass converted in the hydrolysis reaction. We assume that $Q_{out} = \sum_{j=1}^{5} Q_{in,j}$ and thus, the volume of the tank V remains constant throughout the operation of the digester.

3.2 Methanogenesis

In reality, methane fermentation (or methanogenesis) takes multiple steps. Hydrolyzed organic compounds are first fermented to acetate and H_2, then acetate and H_2 are fermented further to CO_2 and CH_4. To simplify our system of equations, we model the entire methane fermentation process from hydrolyzed organic compounds to CH_4 as a single first-order reaction.

We define the chemical oxygen demand (COD) of the homogeneous concentration of hydrolyzed substrate as S_h (in mg / L) and assume that S_h will be consumed during methanogenesis at rate of k_m (in 1/day), a first-order rate constant to approximate overall methane fermentation. Far more precise models of methanogenesis exist, but literature has shown that those models can be simplified considerably when the rate-limiting steps of anaerobic digestion are known [44]. In our case, we assume that hydrolysis or methanogenesis are the two potentially rate-limiting steps, so we believe that combining all fermentation into a single methanogenesis reaction is reasonable.

The rate of change in concentration, dS_h/dt, can be modeled as

$$\frac{dS_h}{dt} = (\sum_j k_{h,j} C_j) - (\frac{Q_{out}}{V} + k_m) S_h \,, \tag{2}$$

wherein, the first term, $\sum_j k_{h,j} C_j$, representing the specific hydrolysis rate of the digester as described by Guo et al. [28]

We rely on stoichiometry to estimate biogas production. There are 0.35 m^3 of CH_4 produced for every kg of COD at standard temperature and pressure [44]. Therefore, we can estimate the daily biogas production, dg/dt, in m^3/day using the equation

$$\frac{dg}{dt} = 0.00035 k_m V S_h / \eta_{CH_4} \tag{3}$$

where η_{CH_4} is the percent of methane in biogas by volume. The factor 0.00035 comes from converting 0.35 m^3 CH_4 / kg COD to 0.00035 L CH_4 / mg COD to account for the units of V (m^3) and S_h (mg / L).

3.3 Sources of Uncertainty

The anaerobic digester receives five distinct waste streams: thickened primary sludge (TPS); thickened waste activated sludge (TWAS); food waste (FW); and fats, oils, and greases (FOG); and scum. As discussed above, each of these streams has a different influent flow rate ($Q_{in,j}$), concentration ($C_{in,j}$), and hydrolysis rate ($k_{h,j}$). Inline flow meters are installed at the facility, so $Q_{in,j}$ is well understood (Figure 1).

The concentrations, $C_{in,j}$, are measured periodically by laboratory technicians, so the data has two potential issues. First, the concentrations vary with time, so this single sample is not representative of the temporal fluctuations in concentration. Second, the single grab sample may not be representative spatially, so that, for a given time it is only an estimate of the waste concentration.

The largest source of uncertainty is the the hydrolysis ($k_{h,j}$) and methanogenesis (k_m) reaction rates. To preserve model simplicity, these reactions are modeled using first-order kinetics. In reality, a complex series of microbial kinetics involving growth and decay of biomass is at play. As a result, these reaction rates vary in time and do not have a true physical meaning (in the way a cell death rate would). Nonetheless, this approach excels at capturing various dimensions of uncertainty from the microbial kinetics in a single reaction term. One final source of uncertainty is the methane fraction of biogas by volume (η_{CH_4}), which varies between 45% and 75% depending on the ratio of acetotrophic to hydrogenotrophic methanogens [31].

Table 1. Parameters in the biogas production model and their range of values from the literature. We assumed an initial concentration of zero for parameters such as food waste concentration that are highly dependent on plant operation and not available from the literature as a result.

Parameter	Meaning	Range (Literature)	Units	Source
k_m	1^{st}-order methanogenesis rate	0.05-0.3	1 / day	[34] [28]
$k_{h,1}$	1^{st}-order TPS hydrolysis rate	0.286-3.0	1 / day	[21] [23]
$k_{h,2}$	1^{st}-order TWAS hydrolysis rate	0.025-0.22	1 / day	[27] [28]
$k_{h,3}$	1^{st}-order FW hydrolysis rate	0.2-0.8	1 / day	[35]
$k_{h,4}$	1^{st}-order FOG hydrolysis rate	0.333-50	1 / day	[5] [30]
$k_{h,5}$	1^{st}-order scum hydrolysis rate	0.1-3.0	1 / day	[21] [23]
η_{CH_4}	Methane fraction of biogas by volume	0.45-0.75		[31]
$C_{1,0}$	Initial concentration of TPS	1,000-50,000	mg / L	[44] [46]
$C_{2,0}$	Initial concentration of TWAS	1,000-50,000	mg / L	[44] [46]
$C_{3,0}$	Initial concentration of FW	Not available		
$C_{4,0}$	Initial concentration of FOG	Not available		
$C_{5,0}$	Initial concentration of scum	Not available		
$S_{h,0}$	Initial hydrolyzed substrate in digester	100-10,000	mg / L	[1] [29] [45]

4 Bayesian Inference

As mentioned earlier in Section 2, we use Bayesian inference to compute intervals over parameters and the initial states, based on some past observations from the system. We assume that the process of biogas production is explained by the model in Section 3, but for unknown parameters and initial conditions. The ranges for these are shown in Table 1. In this section, we describe the overall structure of the probabilistic programming model for performing inference on the biogas production data. We will then briefly summarize the process of extracting posterior samples through a Bayesian inference procedure and the extraction of credible intervals from the samples.

The model used for specifying the prior probabilities $\pi(\theta, \mathbf{x}(0))$ and the generative model $\mathbb{P}(\mathbf{y}(1,\ldots,t) \mid \theta, \mathbf{x}(0), \mathbf{u}(1,\ldots,t))$ is shown in Figure 3. It is a model expressed in the Julia-based Domain Specific Language (DSL) for the state-of-the-art probabilistic programming library `Turing.jl`. The model takes as inputs (Lines 1-6 in the listing of Fig. 3) the input data $\mathbf{u}(1,\ldots,t)$, the output data $\mathbf{y}(1,\ldots,t)$, the ranges for the parameters and initial conditions, as specified in Table 1 and a parameter ξ that specifies the standard deviation for the measurement of the gas production output. Lines 7 - 13 specify the generation of the prior distribution. The function `sim_ode` in line 18 is not shown, but uses an off-the-shelf numerical ODE solver to solve the ODE given parameters, initial conditions and inputs. It then returns the total gas production aggregated for each day in the data. Line 21 conditions that model predicted gas production against the measurements from the data assuming that the measurement device has a known standard error of ξ.

```
@model function gas_production_model(
    inp_data::Matrix{Float64},     # data: inputs to model
    output_data::Vector{Float64},  # data: gas production
    param_ranges::Matrix{Float64}, # parameter ranges
    init_ranges::Matrix{Float64},  # initial condition
    ξ)                             # std. measurement error
    # Prior distribution
    k1 ~ uniform(param_ranges[1,1], param_ranges[1,2])
    k2 ~ uniform(param_ranges[2,1], param_ranges[2,2])
    # ... prior for parameters ...
    # initial conditions
    C10 ~ uniform(init_ranges[1,1], init_ranges[1,2])
    # ... prior for initial conditions ...
    params = [k1, k2, k3, k4, k5, k, eta, V]
    x0 = [C10, C20, C30, C40, C50, S0, G0]
    # Run a ODE simulator given
    #     input data, parameter sample, initial cond. sample
    daily_production = sim_ode(inp_data, params, x0)
    for i in 1:n_days
      # Condition the gas production on the output data
      output_data[i] ~ daily_production[i] + Normal(0.0, ξ)
    end
    return params, init_ranges
```

Fig. 3. Probabilistic programming model in `Turing.jl` for specifying the generative model for the gas production data.

Once the model has been specified, we can use a built-in inference engine in `Turing.jl` to return posterior samples. These samples are of the form

$$S = \{(\theta^{(1)}, \mathbf{x}^{(1)}(0)), \ldots, (\theta^{(N)}, \mathbf{x}^{(N)})\},$$

for a large sample size $N > 0$ (set to $25,000$ for our experiments). The available procedures in `Turing.jl` include Markov chain Monte Carlo (MCMC) methods, Sequential Monte Carlo (SMC) methods and other approaches such as Variational Inference (VI).

Next, given a confidence bound γ, we extract intervals $[\theta_{lo}, \theta_{hi}]$ and $[\mathbf{x}(0)_{lo}, \mathbf{x}(0)_{hi}]$ so that the probability that a given sample $(\theta, \mathbf{x}(t))$ from the posterior distribution belongs to the intervals is $\geq \gamma$.

$$\mathbb{P}(\theta \in [\theta_{lo}, \theta_{hi}] \mid \mathsf{data}) \geq \gamma, \mathbb{P}(\mathbf{x}(0) \in [\mathbf{x}(0)_{lo}, \mathbf{x}(0)_{hi}] \mid \mathsf{data}) \geq \gamma\,.$$

There are many ways of extracting such intervals. Consider the following scheme that works independently with each dimension of θ and $\mathbf{x}(0)$. Consider a dimension z_i of the vector θ or $\mathbf{x}(0)$. We estimate the $z_{i,lo} = \frac{1-\gamma}{2(n+m)}$ quantile and $z_{i,hi} = 1 - \frac{1-\gamma}{2(m+n)}$ quantile of all the values of the scalar z_i from the posterior samples S. Assume that the Bayesian inference procedure satisfies convergence in distribution to the true posterior and that the posterior CDF is continuous everywhere.

Lemma 1. *As the number of samples $N \to \infty$, the posterior probability $\mathbb{P}(z_i \notin [z_{i,lo}, z_{i,hi}]) \to \frac{1-\gamma}{(n+m)}$.*

Proof. This follows from the fact that the empirical distribution $Z_N^{(i)}$, formed by the samples $z_i^{(1)}, \ldots, z_i^{(N)}$ obtained by projecting the samples in set S along z_i, converges to the true posterior $Z^{(i)}$. As a result, the values of $z_{i,lo}$ and $z_{i,hi}$ converge to the $\frac{1-\gamma}{2(n+m)}$ and $1 - \frac{1-\gamma}{2(m+n)}$ quantiles since the CDF is assumed to be continuous as well. Therefore, $\mathbb{P}(z_i \notin [z_{i,lo}, z_{i,hi}]) = \mathbb{P}(z_i \leq z_{i,lo}) + \mathbb{P}(z_i \geq z_{i,hi}) = \frac{1-\gamma}{n+m}$.

The bounds $[\theta_{lo}, \theta_{hi}] \times [\mathbf{x}_{lo}(0), \mathbf{x}_{hi}(0)]$ are obtained as a product over intervals $[z_{i,lo}, z_{i,hi}]$ wherein z_i ranges over the dimensions θ_j for $j = 1, \ldots, m$ and $x_{0,i}$ for $i = 1, \ldots, n$.

Lemma 2. *Assuming the conditions for the convergence in distribution of the Bayesian inference procedure and the continuity of the posterior CDF, as the number of samples $N \to \infty$, the posterior probability $\mathbb{P}((\theta, \mathbf{x}(0)) \notin [\theta_{lo}, \theta_{hi}] \times [\mathbf{x}_{lo}(0), \mathbf{x}_{hi}(0)]) \leq (1-\gamma)$.*

Proof. Proof is obtained by applying the union bound along each dimension to the inequality derived in Lemma 1.

Our approach for deriving the posterior is conservative since it ignores the correlations between the various components of $\theta, \mathbf{x}(0)$ in the posterior. Note that once we have bounds for $[\mathbf{x}_{lo}(0), \mathbf{x}_{hi}(0)]$ at time $t = 0$, we can obtain corresponding bounds at time t $[\mathbf{x}_{lo}(t), \mathbf{x}_{hi}(t)]$ using the reachability analysis algorithm that we will describe in the subsequent section. Finally, we will make the simplifying assumption that $[\theta_{lo}, \theta_{hi}] \subseteq \Theta$. Failing this, we will need to consider the set $[\theta_{lo}, \theta_{hi}] \cap \Theta$, which will be a hyper-rectangle if Θ is a hyper-rectangle.

5 Fast Reachability Analysis using Monotonicity

Thus far, we have used Bayesian inference to obtain posterior bounds in the form of intervals over θ and $\mathbf{x}(0)$. We will now describe a reachability analysis procedure that infers bounds $[\mathbf{x}_{lo}(t), \mathbf{x}_{hi}(t)]$ for the states at time t and therefore results in output bounds $[\mathbf{y}_{lo}(t), \mathbf{y}_{hi}(t)]$. There are many reachability analysis tools that can be used off-the-shelf for such analysis including Flow* [18] and CORA [2]. Note that although the model presented in Section 3 is linear in the state-variables, the uncertainties in the parameters θ means that standard approaches to linear systems that have been shown to work for billions of state variables cannot be directly applied [6]. In this section, we present an efficient approach by exploiting some key physical properties of the wastewater model and ideas from the study of positive differential equations. Our approach is based on ideas presented by Meyer et al [40]. The key contributions include showing that the WWTP model in Section 3 satisfies the necessary monotonicity conditions; and in doing so, adapting the ideas of Meyer et al to systems with uncertain parameters.

The ODE model in Section 3 has the form $\frac{d\mathbf{x}}{dt} = A(\theta)\mathbf{x} + h(\mathbf{u})$, wherein $A(\theta)$ is a matrix whose entries are affine functions (linear plus constant) over the parameters θ and $h(\mathbf{u})$ is an input dependent term that is always non-negative.

Definition 2 (Metzler Matrix). *A matrix $M \in \mathbb{R}^{n\times n}$ is said to be* Metzler *(essentially positive) iff $M_{i,j} \geq 0$ for $i \neq j$.*

An ODE is positive iff whenever $\mathbf{x}(0) \geq 0$ then $\mathbf{x}(t) \geq 0$ for all time $t \geq 0$. The notion of positivity is significant because the waste-water treatment plant involves physical quantities such as concentration of effluents that are non-negative at all times (negative concentrations are not physically meaningful).

Theorem 3. *Consider the ODE $\frac{d\mathbf{x}}{dt} = M\mathbf{x} + \mathbf{b}$ for vector $\mathbf{b} \geq 0$ entrywise. The system is positive if and only if M is a Metzler matrix.*

Proof can be found in a standard textbook [24].

Lemma 3. *The model $\frac{d\mathbf{x}}{dt} = A(\theta)\mathbf{x} + h(\mathbf{u})$ defined in Section 3 satisfies the condition that $A(\theta)$ is Metzler for all $\theta \in [\theta_{lo}, \theta_{hi}]$ and $h(\mathbf{u}) \geq 0$ for all $\mathbf{u} \in U$.*

Proof is by verifying each entry $A_{i,j}(\theta) \geq 0$ for $i \neq j$ and $\theta \in [\theta_{lo}, \theta_{hi}]$. From Eq. 1, the off-diagonal terms include $\frac{Q_{in,j}}{V} \geq 0$ and from Eq. (2), the off-diagonal terms are $k_{h,j} \geq 0$ (see Table 1).

Lemma 4. *For any $\theta \in [\theta_{lo}, \theta_{hi}]$ such that $\theta_{hi} \geq \theta_{lo} \geq 0$, the matrix $A(\theta)$ for the model in Section 3 is Metzler. Also, there exists Metzler matrices L, U such that*

$$\forall\, \theta \in [\theta_{lo}, \theta_{hi}],\ L \leq A(\theta) \leq U,$$

wherein the inequality $\leq$ between matrices is interpreted entrywise.

Proof. The matrices L, U are obtained by computing for each entry in $A(\theta)$, $L_{i,j} = \min_{\theta\in[\theta_{lo},\theta_{hi}]} A_{i,j}(\theta)$ and $U_{i,j} = \max_{\theta\in[\theta_{lo},\theta_{hi}]} A_{i,j}(\theta)$. Since $[\theta_{lo}, \theta_{hi}]$ form a compact interval, it is easy to see that the minima and maxima exist. Furthermore, L is Metzler following Lemma 3 for each $i \neq j$, the minimum value of $A_{i,j}(\theta) \geq 0$. Since $L \leq U$ entrywise, U is Metzler as well.

Consider the time trajectories $\varphi_{lo}(t)$ of the initial value problem (IVP) $\frac{d\mathbf{x}_{lo}}{dt} = L\mathbf{x}_{lo} + h(\mathbf{u})$ for initial condition $\mathbf{x}_{lo}(0)$ and $\varphi_{hi}(t)$ for the IVP $\frac{d\mathbf{x}_{hi}}{dt} = U\mathbf{x}_{hi} + h(\mathbf{u})$ with initial condition $\mathbf{x}_{hi}(0)$.

Theorem 4. *For all $\theta \in [\theta_{lo}, \theta_{hi}]$, $\mathbf{x}(0) \in [\mathbf{x}_{lo}, \mathbf{x}_{hi}]$ and time $t \geq 0$, the solution $\psi(t)$ of the system $\frac{d\mathbf{x}}{dt} = A(\theta)\mathbf{x} + h(\mathbf{u})$ satisfies the bounds $\varphi_{lo}(t) \leq \psi(t) \leq \varphi_{hi}(t)$.*

Note that φ_{lo} is the solution of the system $\frac{d\mathbf{x}_{lo}}{dt} = L\mathbf{x}_{lo} + h(\mathbf{u}(t))$ with initial condition $\mathbf{x}_{lo}(0)$ while φ_{hi} is the solution of the system $\frac{d\mathbf{x}_{hi}}{dt} = U\mathbf{x}_{hi} + h(\mathbf{u}(t))$ with initial condition $\mathbf{x}_{hi}(0)$. For any fixed θ, let $M = A(\theta)$ be a Metzler matrix (by assumption).

Lemma 5. *If $L \leq M$ then for all $\mathbf{x} \geq 0$, $L\mathbf{x} \leq M\mathbf{x}$.*

Proof. Note that the matrix $M - L$ has all non-negative entries since $L \leq M$. Therefore, $(M - L)\mathbf{x} \geq 0$ whenever $\mathbf{x} \geq 0$. Thus, $L\mathbf{x} \leq M\mathbf{x}$.

We recall a well-known result known as the monotone comparison principle, that is also sometimes known as the Chaplygin's theorem. We specialize this for the case of linear systems in our presentation below.

Theorem 5 (Monotone Comparison Principle). *Consider a dynamical system* $\frac{d\mathbf{x}}{dt} = M\mathbf{x} + h(\mathbf{u}(t))$*, wherein* M *is Metzler and* $h(\mathbf{u}) \geq 0$ *for all* $\mathbf{u}$*. Consider a system* $\frac{d\mathbf{z}}{dt} = L\mathbf{z} + h(\mathbf{u}(t))$*, wherein* $L \leq M$*. If* $\mathbf{z}(0) \leq \mathbf{x}(0)$ *at time* $t = 0$*, then* $\mathbf{z}(t) \leq \mathbf{x}(t)$ *for all time.*

Note that L, M are both Metzler matrices. Furthermore, $\mathbf{x}_{lo}(0) \leq \mathbf{x}(0)$. As a result, applying the monotone comparison principle, yields the result that $\mathbf{x}_{lo}(t) \leq \mathbf{x}(t)$. Similarly, M, U are Metzler matrices and $\mathbf{x}(0) \leq \mathbf{x}_{hi}(0)$. Applying the theorem yields the result that $\mathbf{x}(t) \leq \mathbf{x}_{hi}(t)$. This concludes the proof of Theorem 4.

Theorem 4 is remarkable since it reduces the reachability analysis problem to that of computing the solutions of two linear systems for fixed matrices and initial conditions. We perform the reachability analysis as follows:

1. Compute the matrices L, U that form upper and lower bounds of $A(\theta)$.
2. Compute a rigorous lower-bound $\varphi_{lo}(t)$ for the system $\frac{d\mathbf{x}_{lo}}{dt} = L\mathbf{x}_{lo} + h(\mathbf{u})$ with initial condition $\mathbf{x}_{lo}(0)$.
3. Compute a rigorous upper-bound $\varphi_{hi}(t)$ for the system $\frac{d\mathbf{x}_{hi}}{dt} = U\mathbf{x}_{hi} + h(\mathbf{u})$ with initial condition $\mathbf{x}_{hi}(0)$.
4. The reachable set at time $t \geq 0$ is contained in $[\varphi_{lo}(t), \varphi_{hi}(t)]$.

6 Implementation and Data Sources

We have implemented both the Bayesian inference and flowpipe construction in Julia. The historical data used to evaluate our algorithm comes from a municipal WWTP in the San Francisco Bay Area (Cf. Section 6.1). The ability for our algorithm to perform on real-world data is vital given our industrial application.

6.1 Data Sources

We use historical data from our case study WWTP to evaluate our method. This data includes a variety of measurements from different sources and on different timescales. For example, flow rates of the various waste streams are recorded in close to real time by inline sensors, while the strength of each waste stream (i.e., concentration) is measured in a laboratory every few days. We perform our analysis on daily timesteps due to the different granularity of data. I.e., we sum or average the sub-daily flow rate data to daily values and linearly interpolate the concentration data between missing days so that all the data is on the same timescale. We do not clean the data besides interpolating missing values, so the typical noise and potential inaccuracies of real-world operational data are present in our dataset. Our combined method of Bayesian inference followed by flowpipe construction is especially suited to real-world data with large uncertainty such as this case study. Further description of specific data used are found in the Appendix (available upon request).

We use data from three sample months for our analysis: September 2018, January 2020, and March 2020. These months include different operating conditions, such as influent waste stream concentrations and reaction rate parameters, to demonstrate the generalizability of our method at each step.

6.2 Bayesian Inference

The first step of our two-step approach is Bayesian inference to determine the credible intervals of unknown parameters. We perform this Bayesian inference using an MCMC method (we use 25,000 iterations of the Metropolis Hastings algorithm) from the `Turing.jl` package [25]. In general, selecting the number of iterations for Bayesian inference is hard. We chose to perform 25,000 iterations after initial tuning experiments that compared the variability of the confidence intervals obtained after each run versus the time taken to run. For $25,000$ iterations, the results had no variation across multiple runs. We initialize the algorithm with the nominal parameters and initial values of concentrations that are available in the Appendix (available upon request).

We chunk the historical data into 8-day segments, wherein the first 3 days of data are used for Bayesian inference to determine the kinetic parameters and initial conditions of the wastewater model. We then construct an 8-day flowpipe from the beginning of the Bayesian inference, rejecting the first 3 days that overlap with the training period to produce a 5-day prediction. The prediction uses the ground truth inputs for the next 5 days but in an application, it may consist of a future scenario that the plant operators may consider for their analysis. We also use these samples to construct posterior predictive simulations for the next 5 days using an ODE solver and compare the flowpipe against the posterior predictive simulations. In Section 7, we evaluate our approach for 75%, 90%, and 99% credible intervals.

6.3 Flowpipe Construction

As mentioned in Section 5, we can leverage the monotonicity of the wastewater model presented in Section 3 to quickly over-approximate the bounds of biogas production. We begin the flowpipe construction by taking the parameter estimates from the Bayesian inference for the desired credible interval. Once we define the initial conditions using the output from the Bayesian inference, we simply propagate those bounds from the same start date as the Bayesian inference. We perform reachability analysis for a 8-day period, then ignoring the 3 days that overlap with the Bayesian inference, we evaluate the flowpipe on the final 5 days of the simulation period. Specific implementation details in Julia are described in the Appendix (available upon request).

7 Evaluation

We evaluate our two-step approach of Bayesian inference and flowpipe construction versus the standard approach of performing posterior predictive simulations. We compare the accuracy and computational time of our predictions for credible intervals (CI) of 75%, 90%, and 99%. The code used for evaluation can be found in the software artifact available on Zenodo [14].

Posterior Predictive Simulations: We compare our approach against a standard approach based on constructing the so-called "posterior predictive distribution" [38] through simulations. The posterior predictive simulations approach uses the samples computed through Bayesian inference to predict future states of the model through simulation. In our implementation, we simulate the samples computed by the Bayesian inference procedure using the ODE model to obtain sample values for gas production for the 5 days in the future. We then simply compute credible intervals for each of the future days as an alternative to flowpipe construction. Note that the simulation is performed for each sample obtained by the Bayesian inference algorithm.

In Figure 4, we illustrate representative prediction outcomes across different time periods. Table 2 summarizes statistics across the entire data set spanning three months split into multiple 8 day segments. Both the two-step approach (Figure 4, left) and the approach using posterior predictive simulations (Figure 4, right) provide reasonable prediction bounds that capture the historical trajectory. Although the posterior predictive simulations produce tighter bounds, they fail to accurately predict the ground truth. Whereas, the approach proposed here using flowpipe construction is able to capture the ground truth data with very high accuracy.

We hypothesize that the use of credible intervals captures a simple but accurate set of point estimates from the posterior samples. The rest of the approach is entirely free of any randomness and therefore has generalized from the available samples to facts about the underlying probability distributions. On the other hand, the posterior simulations are tied to the samples and attempt to make complex inference (by propagating the samples through a differential equation model) that fails to generalize well. The high uncertainty in measurements and changing conditions of the wastewater treatment processes exacerbate the challenges faced by the posterior simulation approach.

Table 2 presents key performance statistics comparing the two methods. The two-step approach consistently outperforms Bayesian inference alone in terms of prediction *accuracy*, defined as the proportion of ground-truth gas production values captured within the predicted bounds. For example, the combined method captures 94% of samples at the 90% CI, while the posterior simulations method captures only 21% of the samples.

Table 2. Performance statistic comparing our two-step approach of Bayesian inference followed by flowpipe construction to posterior predictive simulations approach for various credible intervals (CIs). Legend: **Accuracy** refers to the percentage of samples where the historical (ground-truth) biogas production fell within the predicted bounds; **Uncertainty** refers to the average width of the bounds in m^3; **Bayesian Inference Time** refers to the amount of time it took to perform the 3-day initialization, and **Prediction Time** refers to the 8-day prediction. All computation times were gathered on a 2023 MacBook Pro using the Apple M2 Pro chip with 12 cores.

	Flowpipe Construction			Posterior Predictive Simulations		
	75% CI	90% CI	99% CI	75% CI	90% CI	99% CI
Accuracy	0.85	0.94	1.00	0.13	0.21	0.33
Uncertainty (m^3)	9220	13850	21290	330	490	770
Bayesian Inference Time (s)	64	64	64	64	64	64
Prediction time (s)	1.6	1.7	1.6	77	77	77

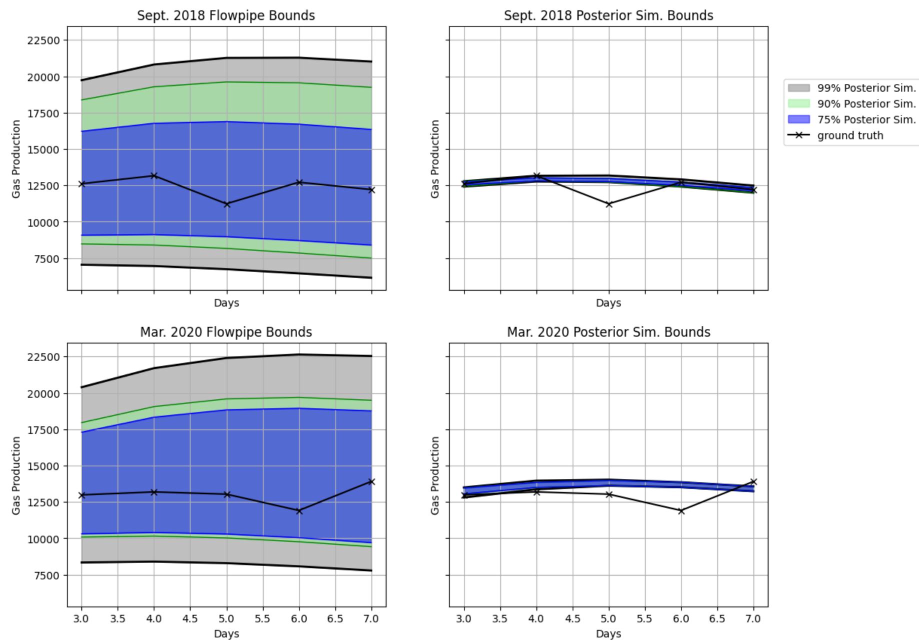


Fig. 4. Sample predictions for a single 8 day segment from September 2018 (top) and March 2020 (bottom) using various methods for various credible intervals (CIs). (Left) Upper and lower bounds from our two-step approach of Bayesian inference and flowpipe construction. (Right) Credible intervals from posterior prediction.

This improved accuracy comes with increased *uncertainty*, i.e., wider bounds. At the 99% CI, the combined method produces an average interval width of 21,290 m^3. However, the fact that the posterior predictive samples have a small credible interval is irrelevant since their accuracy is quite low.

Importantly, the *inference time* remains constant across both methods and all CIs (64 seconds), since both use the same Bayesian inference stage.

The most notable difference lies in *prediction time*: whereas posterior simulations takes over 75 seconds, the combined method completes in under 2 seconds, regardless of CI level. This $40\times$ speedup enables the method to be used in time-sensitive applications without sacrificing reliability.

To conclude, our approach based on credible intervals and flowpipe computations yields not just a speedup but also captures the ground truth data well when compared to the standard approach of posterior predictive simulations.

8 Conclusion and Future Work

In this paper, we present a novel combination of two heavily researched mathematical topics: probabilistic programming and reachability analysis. Specifically, we use MCMC on 3-days of past data to create credible intervals for the unknown parameters and initial conditions of a first-order kinetics model of biogas production at a WWTP.

We then exploit the monotonicity of this model to quickly compute upper and lower bounds for the next 5-days of biogas production at the WWTP. These formal bounds could be used to verify biogas production predictions at WWTPs, which in turn facilitates energy management plans.

We found it challenging to evaluate our results against previous publications given the dearth of prior research in this area. Instead, we perform posterior predictive simulations for the same 5-day period that we construct a flowpipe using reachability analysis. Across credible intervals from 75% to 99%, we found that posterior predictive simulations had an accuracy of 13-31% compared to 85-100% accuracy for our two-step Bayesian inference plus flowpipe construction. High accuracy is vital to building the trust of operators when deploying prediction algorithms in the context of automated control, especially for critical infrastructure like WWTPs. The much faster computation time of flowpipe construction versus posterior predictive simulations (approximately 40x in the prediction phase) makes it better suited for automated planning and closed-loop controls.

Our method seeks to construct formal bounds given the sensitive nature of WWTPs as critical infrastructure. We are successful, as we achieve a 100% accuracy for a 99% credible interval, but this comes with a large degree of uncertainty (21,290 m^3 for a 99% credible interval). Unfortunately, this large uncertainty makes the algorithm impractical to deploy in a real-world setting. The 75% credible interval is still quite accurate with the biogas production falling within the bounds 85% of the time, and it has less than half the uncertainty of the 99% credible interval.

More credible intervals could be tested, and the operators could give input on the balance of accuracy and uncertainty that their prefer at their WWTP. Future work will also study the effect of using inference algorithms other than MCMC such as sequential Monte Carlo and other methods. With further refinement, our combination of probabilistic programming and reachability analysis has the potential to verify predictions of biogas production at WWTPs. These predictions of biogas production are vital to successfully deploying cost- or emissions-minimizing operational strategies that have been recently studied at WWTPs [11, 16]. One challenge to deploying these optimal energy management plans is the role of WWTPs as critical infrastructure and subsequent risk aversion of operators, so formal methods offer a compelling approach to reassuring WWTP operators of the safety of these operational recommendations. This technique could also be applied to critical infrastructure outside of the water sector that can be modeled with ODEs, such as transportation and energy.

Acknowledgments

This work was funded, in part, by the Center for Integrated Facility Engineering at Stanford University as a part of CIFE Seed Proposal 2023-02, "Formal proofs of safe operating limits at wastewater resource recovery facilities"; the Stanford Woods Institute for the Environment as a part of Realizing Environmental Innovation Program (REIP) under award “Affordable, Sustainable, and Just Design of Next Generation Wastewater Resource Recovery Facilities”; and the US National Science Foundation (NSF) under awards # CCF-2422136 and CPS-1836900.

References

1. Ahnert, M., Schalk, T., Brückner, H., Effenberger, J., Kuehn, V., Krebs, P.: Organic matter parameters in wwtp–a critical review and recommendations for application in activated sludge modelling. Water Science and Technology **84**(9), 2093–2112 (2021)
2. Althoff, M.: An introduction to CORA 2015. In: Proc. Applied Verification for Continuous and Hybrid Systems (ARCH). pp. 120–151. EasyChair (December 2015). https://doi.org/10.29007/zbkv, https://easychair.org/publications/paper/xMm
3. Althoff, M., Dolan, J.M.: Online verification of automated road vehicles using reachability analysis. IEEE Transactions on Robotics **30**(4), 903–918 (2014)
4. Althoff, M., Frehse, G., Girard, A.: Set propagation techniques for reachability analysis. Annual Review of Control, Robotics, and Autonomous Systems **4**(1), 369–395 (2021). https://doi.org/10.1146/annurev-control-071420-081941
5. Baena, A., Orjuela, A., Rakshit, S.K., Clark, J.H.: Enzymatic hydrolysis of waste fats, oils and greases (fogs): Status, prospective, and process intensification alternatives. Chemical Engineering and Processing-Process Intensification **175**, 108930 (2022)
6. Bak, S., Tran, H.D., Johnson, T.T.: Numerical verification of affine systems with up to a billion dimensions. In: Proceedings of the 22nd ACM International Conference on Hybrid Systems: Computation and Control. pp. 23 – 32. HSCC '19, Association for Computing Machinery, New York, NY, USA (2019)
7. Bansal, S., Bajcsy, A., Ratner, E., Dragan, A.D., Tomlin, C.J.: A hamilton-jacobi reachability-based framework for predicting and analyzing human motion for safe planning. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 7149–7155. IEEE (2020)
8. Barthe, G., Katoen, J.P., Silva, A.: Foundations of Probabilistic Programming. Cambridge University Press (2020)
9. Berz, M., Makino, K.: Verified integration of odes and flows using differential algebraic methods on high-order taylor models. Reliable computing **4**(4), 361–369 (1998)
10. Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D.: Pyro: deep universal probabilistic programming. J. Mach. Learn. Res. **20**(1), 973–978 (Jan 2019)
11. Bolorinos, J., Mauter, M.S., Rajagopal, R.: Integrated energy flexibility management at wastewater treatment facilities. Environmental Science & Technology **57**(46), 18362–18371 (2023)
12. Cairoli, F., Paoletti, N., Bortolussi, L.: Conformal quantitative predictive monitoring of stl requirements for stochastic processes. In: Proceedings of the 26th ACM International Conference on Hybrid Systems: Computation and Control. pp. 1–11 (2023)
13. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A probabilistic programming language. Journal of statistical software **76**(1) (2017)
14. Chapin, F., Varma, A., Akinwande, S., Mauter, M., Sankaranarayanan, S.: Software artifact for using bayesian inference and flowpipe construction to bound predictions of biogas production at wastewater treatment plants (Aug 2025). https://doi.org/10.5281/zenodo.16930822
15. Chapin, F.T., Bolorinos, J., Mauter, M.S.: Electricity and natural gas tariffs at united states wastewater treatment plants. Scientific Data **11**(1), 113 (2024)
16. Chapin, F.T., Wettermark, D., Bolorinos, J., Mauter, M.S.: Load-shifting strategies for cost-effective emission reductions at wastewater facilities. Environmental Science & Technology (2025)

17. Chen, X., Abraham, E., Sankaranarayanan, S.: Taylor model flowpipe construction for non-linear hybrid systems. In: 2012 IEEE 33rd Real-Time Systems Symposium. pp. 183–192. IEEE (2012)
18. Chen, X., Ábrahám, E., Sankaranarayanan, S.: Flow*: An analyzer for non-linear hybrid systems. In: Computer Aided Verification: 25th International Conference, CAV 2013, Saint Petersburg, Russia, July 13-19, 2013. Proceedings 25. pp. 258–263. Springer (2013)
19. Chen, X., Sankaranarayanan, S.: Reachability analysis for cyber-physical systems: Are we there yet? (invited paper). In: Proc. NASA Formal Methods Symposium. Lecture Notes in Computer Science, vol. 13260, pp. 109–130. Springer (2022)
20. Chou, Y., Yoon, H., Sankaranarayanan, S.: Predictive runtime monitoring of vehicle models using bayesian estimation and reachability analysis. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2111–2118. IEEE (2020)
21. Christensen, M.L., Jakobsen, A.H., Hansen, C.S.K., Skovbjerg, M., Andersen, R.B.M., Jensen, M.D., Sundmark, K.: Pilot-scale hydrolysis of primary sludge for production of easily degradable carbon to treat biological wastewater or produce biogas. Science of the Total Environment **846**, 157532 (2022)
22. Clarke, E.M., Zuliani, P.: Statistical model checking for cyber-physical systems. In: International symposium on automated technology for verification and analysis. pp. 1–12. Springer (2011)
23. Eastman, J.A., Ferguson, J.F.: Solubilization of particulate organic carbon during the acid phase of anaerobic digestion. Journal (Water Pollution Control Federation) pp. 352–366 (1981)
24. Farina, L., Rinaldi, S.: Positive Linear Systems: Theory and Applications. John Wiley & Sons, Inc. (2000), https://onlinelibrary.wiley.com/doi/book/10.1002/9781118033029
25. Ge, H., Xu, K., Ghahramani, Z.: Turing: a language for flexible probabilistic inference. In: International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain. pp. 1682–1690 (2018), http://proceedings.mlr.press/v84/ge18b.html
26. Ghosh, B., Duggirala, P.S.: Reachability of linear uncertain systems: Sampling based approaches. arXiv preprint arXiv:2109.07638 (2021)
27. Gossett, J.M., Belser, R.L.: Anaerobic digestion of waste activated sludge. Journal of the Environmental Engineering Division **108**(6), 1101–1120 (1982)
28. Guo, H., Oosterkamp, M.J., Tonin, F., Hendriks, A., Nair, R., van Lier, J.B., de Kreuk, M.: Reconsidering hydrolysis kinetics for anaerobic digestion of waste activated sludge applying cascade reactors with ultra-short residence times. Water Research **202**, 117398 (2021)
29. Henze, M., van Loosdrecht, M.C., Ekama, G.A., Brdjanovic, D.: Biological wastewater treatment. IWA publishing (2008)
30. Iasmin, M., Dean, L.O., Ducoste, J.J.: Quantifying fat, oil, and grease deposit formation kinetics. Water Research **88**, 786–795 (2016)
31. International Energy Agency: Outlook for biogas and biomethane: prospects for organic growth. OECD Publishing, Paris (2020)
32. Jaynes, E.T.: Confidence intervals vs bayesian intervals. In: Harper, W.L., Hooker, C.A. (eds.) Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, pp. 175–. D. Reidel, Dordrecht (1976)
33. Jha, S.K., Clarke, E.M., Langmead, C.J., Legay, A., Platzer, A., Zuliani, P.: A bayesian approach to model checking biological systems. In: Computational Methods in Systems Biology: 7th International Conference, CMSB 2009, Bologna, Italy, August 31-September 1, 2009. Proceedings 7. pp. 218–234. Springer (2009)
34. Koch, K., Drewes, J.E.: Alternative approach to estimate the hydrolysis rate constant of particulate material from batch data. Applied Energy **120**, 11–15 (2014)

35. Koch, K., Helmreich, B., Drewes, J.E.: Co-digestion of food waste in municipal wastewater treatment plants: Effect of different mixtures on methane yield and hydrolysis rate constant. Applied Energy **137**, 250–255 (2015)
36. Lal, R., Prabhakar, P.: Bounded error flowpipe computation of parameterized linear systems. In: 2015 International Conference on Embedded Software (EMSOFT). pp. 237–246. IEEE (2015)
37. Mahmood, Q., Ping, Z., Li, G.X., Mei, L.L.: The rate-limiting step in anaerobic digestion in the presence of phosphine. Toxicology and industrial health **22**(4), 165–172 (2006)
38. McElreath, R.: Statistical Rethinking: A Bayesian Course with Examples in R and STAN. Chapman & Hall/CRC Texts in Statistical Science, Boca Raton, FL, 2nd edn. (2020)
39. van de Meent, J.W., Paige, B., Yang, H., Wood, F.: An introduction to probabilistic programming (2021), https://arxiv.org/abs/1809.10756
40. Meyer, P.J., Coogan, S., Arcak, M.: Sampled-data reachability analysis using sensitivity and mixed-monotonicity. IEEE Control Systems Letters **2**(4), 761–766 (2018). https://doi.org/10.1109/LCSYS.2018.2848280
41. Musabandesu, E., Loge, F.: Load shifting at wastewater treatment plants: A case study for participating as an energy demand resource. Journal of Cleaner Production **282**, 124454 (2021)
42. Narayanan, P., Carette, J., Romano, W., Shan, C., Zinkov, R.: Probabilistic inference by program transformation in hakaru (system description). In: Kiselyov, O., King, A. (eds.) Functional and Logic Programming. FLOPS 2016. Lecture Notes in Computer Science, vol. 9613, pp. 1–29. Springer, Cham (2016)
43. Reifsnyder, S., Cecconi, F., Rosso, D.: Dynamic load shifting for the abatement of ghg emissions, power demand, energy use, and costs in metropolitan hybrid wastewater treatment systems. Water Research **200**, 117224 (2021)
44. Rittmann, B.E., McCarty, P.L.: Environmental Biotechnology: Principles and Applications. McGraw-Hill Education, New York, second edn. (2020)
45. Roeleveld, P., Van Loosdrecht, M.: Experience with guidelines for wastewater characterisation in the netherlands. Water Science and Technology **45**(6), 77–87 (2002)
46. Sillanpää, M., Khadir, A., Gurung, K.: Resource Recovery in Drinking Water Treatment. Elsevier (2023)
47. Sowby, R.B.: Increasing water and wastewater utility participation in demand response programs. Energy Nexus **1**, 100001 (2021)
48. Tolpin, D., van de Meent, J.W., Wood, F.: Probabilistic programming in anglican. In: Proceedings of the 2015th European Conference on Machine Learning and Knowledge Discovery in Databases. p. 308–311. ECMLPKDD'15, Springer (2015)
49. Vavilin, V., Fernandez, B., Palatsi, J., Flotats, X.: Hydrolysis kinetics in anaerobic degradation of particulate organic material: an overview. Waste management **28**(6), 939–951 (2008)
50. Zohrabian, A., Plata, S.L., Kim, D.M., Childress, A.E., Sanders, K.T.: Leveraging the water-energy nexus to derive benefits for the electric grid through demand-side management in the water supply and wastewater sectors. Wiley Interdisciplinary Reviews: Water **8**(3), e1510 (2021)
51. Zuliani, P.: Statistical model checking for biological applications. Int. J. Softw. Tools Technol. Transf. **17**(4), 527–536 (Aug 2015)