DEVELOPMENT OF A HYBRID MODEL AND OPTIMAL CONTROL ALGORITHM FOR A FULL-SCALE BIO-FERMENTATION PROCESS

A Thesis

by

PARTH JITENDRA SHAH

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Dr. Joseph Kwon
Committee Members,	Dr. Costas Kravaris
	Dr. Eduardo Gildin
Head of Department,	Dr. Arul Jayaraman

December 2021

Major Subject: Chemical Engineering

ABSTRACT

Bio-fermentation process is difficult to model given its use of living micro-organisms to produce useful products via complex reaction mechanisms. Their kinetics are hard to characterize; hence, approximate formulations are used when building a first-principles model. Consequently, such a model will be of poor accuracy. Recently, there is a lot of interest towards data-driven modeling as the amount of data collected, stored, and utilized is growing tremendously due to the advent of super-computing power and data storage device. Additionally, data-driven models are simple and easy to build but their utility is hugely restricted by the amount and quality of data used to develop them. Therefore, hybrid modeling is an attractive alternative to purely data-based modeling, wherein it combines a first-principles model with a data-based model resulting in improved accuracy and robustness. In this work, we develop a three-step method to build a hybrid model for a full-scale bio-fermentation process with a volume of over 100,000 gallons. Firstly, we improved the accuracy of the first-principles model via incorporating mathematical terms in its equations which are based on obtained process knowledge from a literature study. Secondly, we performed local and global sensitivity analysis to identify sensitive parameters in the improved first-principles model that have considerable influence on its prediction capability. Finally, we developed a deep neural network (DNN) based hybrid model by integrating the improved first-principles model with a DNN which is trained to predict the identified model parameters. The resulting hybrid model is more accurate and robust than the (original and improved) first-principles models as it is equipped with a trained DNN to predict the uncertain parameters and process states accurately.

Based on the developed hybrid model, a hybrid model-based observer was developed to track the different states present in the process. As the available measurements were fairly accurate, the open-loop observer was re-initialized with a new set of measurements whenever they become available. This method is computationally less demanding and was able to accurately estimate the states. Next, we build an optimal control algorithm on GAMS software to estimate the optimal operating conditions of the fermenter in real-time. This is carried out in order to maximize

the product amount and minimize the cost by manipulating the inputs and taking practical constraints into account. The resulting control algorithm was able to improve the profitability and the productivity of the full-scale bio-fermentation process.

DEDICATION

To my mother, father, and grandparents.

ACKNOWLEDGMENTS

Firstly, I would like to take this opportunity to express my utmost gratitude and respect to my advisor Dr. Joseph Sang Il-Kwon for his endless support and motivation throughout my Master's. His encouragement and supervision inspired to work hard and be engaged in my research. I would also like to express my appreciation and thank him for his continued guidance, ideas, discussions, and thoughtful comments. I will always treasure all that I have I have learned from him and would like to heartily thank him for his dedication and support. I would also like to appreciate and thank my committee members, Dr. Costas Kravaris and Dr. Eduardo Gildin for all their assistance and timely help. A special thanks to Dr. Kravaris for his valuable ideas, discussions and help throughout this work.

I would like to give my greatest appreciation to Dr. Ziyan Sheriff and Mr. Mohammed Saad Faizan Bangi for their continued help and collaboration. I sincerely thank both of them as this work would have not been possible without their ideas, fruitful discussions, and constructive feedback. I also thank my research colleagues and friends Dr. Prashanth Siddhamshetty, Dr. Abhinav Narasingam, Dr. Dongheon Lee, Mr. Hyun-Kyu Choi, Ms. Pallavi Kumari, Ms. Bhavana Bhadriraju, Mr. Kaiyu Cao, Mr. Niranjan Sitapure, Mr. Silabrata Pahari, and all visiting students, for all their help and support both within and outside the research group. It is an enriching and memorable experience for me to work with all of them in our research group.

A very special thanks to all my friends at Texas A&M, especially Haard, Kaushal, Anubhav, Ratul, and Suyash. Thank you for always being there for me and sharing this journey. I am also thankful to my beloved friends Srikanth, Nirabhra, Simran, and Vatsalya for all their care and encouragement.

Last but not the least, I owe my heartfelt gratitude to my parents, grandparents, uncles, aunts, and cousins for all their love and care. I am forever indepted to my family and thank them for their daily support and motivation.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Dr. Joseph Sang-Il Kwon [principal advisor] of the Department of Chemical Engineering and Texas A&M Energy Institute, Dr. Costas Kravaris of the Department of Chemical Engineering, and Dr. Eduardo Gildin of the Department of Petroleum Engineering.

All the work conducted for the thesis was completed by the student independently in a collaboration with industrial sponsor Kaneka Corporation, Dr. Costas Kravaris, Dr. Ziyan Sheriff, and Mr. Mohammed Saad Faizan Bangi.

Funding Sources

The authors gratefully acknowledge financial support from Kaneka Corporation, the Texas A&M Energy Institute, and the Artie McFerrin Department of Chemical Engineering.

NOMENCLATURE

DNN	Deep Neural Network
ML	Machine Learning
MPC	Model Predictive Control
GAMS	General Algebraic Modeling System software
ReLu	Rectified Linear Unit
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
SSE	Sum of Squared Error
RMSE	Root Mean Squared Error
FIM	FIsher Information Matrix
ϕ_D	D-optimality criterion
$lhsdesign$	Latin Hypercube Sampling
RE	Relative Error
NN	Neural Network
NLP	Non-linear programming
EKF	Extended Kalman Filter
UKF	Unscented Kalman Filter

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Bio-fermentation processes are widely used for industrial production of many useful products such as chemicals, enzymes, food products, and pharmaceuticals. They involve the use of micro-organisms as 'catalysts' which convert substrates to products of interest. These micro-organisms can be bacteria, fungi, mammalian cells, etc., and are often optimized and engineered to achieve greater yields of product than observed in naturally occurring systems. These processes are advantageous over chemical processes as they are sustainable due to there low-temperature and low-pressure operations, and no requirements for harsh chemicals (1). A typical bio-fermentation process is carried out in two phases. During the first phase, a bulk of substrate is combined with micro-organisms and other essential nutrients which are required for their growth. During this phase, the micro-organisms consume the available substrate, and consequently, there is an increase in the biomass concentration. In the second phase, other additional substrates are continuously fed into the reactor, and the rates of feeding are heavily regulated to avoid overfeeding or underfeeding which can significantly reduce the productivity of the process. Product is recovered from the reactor either continuously during the process or at the end of the second phase.

Modeling a bio-fermentation process is a challenging task given the complex interactions that occur within it. Usually, a first-principles model is developed using mass and energy conservation laws, kinetic laws, thermodynamic laws, etc., and it is able to capture the essential dynamics of the process. For this reason, building such a model requires significant time, resources, and process insight. Additionally, due to the complex nature of the process, some mechanisms within the process are not understood to a level that they can be accurately modeled, and in such cases, empirical formulations are introduced in the first-principles model. The overall accuracy of the first-principles model is dependent on these empirical relationships. On the other hand, a data-based model can be developed using historical process data, which is easy to build and is accurate in its training regime, but will not be robust over a wide range of operating conditions of the process (2). Another class of models called hybrid models exist which are a combination of first-

principles models and data-based models (3).

In process modeling, the idea of hybrid models was advanced from the field of neural networks (4), which was to build hybrid models by combining NNs with first-principles knowledge. This combination led to models with higher accuracy than the first-principles model, and better extrapolation and interpretation capabilities than solely NN-based black-box models. Now, there are many kinds of hybrid models depending on the nature of combination between first-principles models and data-based models. For instance, during the 1990s, the concept of grey-box models was used in systems and control theory where structural information from first-principles models was incorporated into the black-box models (5). But, since then, the understanding of the term 'hybrid model' has grown to represent any combination of a first-principles model with a data-based model, and it has been applied in various chemical and biochemical engineering applications, which includes: modeling of bacteria cultivations (6), chemical reactor (7; 8), crystallization (9), distillation columns (10), fungi cultivations (11), hybridoma cell cultivations, (12), insect cell cultivations (13), mammalian cell cultivations (14), mechanical reactors (15), metallurgic processes (16; 17), milling (18), polymerization processes (19), thermal devices (20), yeast fermentations (21; 22), hydraulic fracturing (23), intracellular signaling pathway (24), etc. For more information on hybrid modeling in the field of process systems engineering, one can view (25), an excellent review paper.

Given the advantages of a hybrid model, it has been widely used to model lab-scale or pilot-scale bio-fermentation processes. Hybrid models usually involve an ANN as a function approximator predicting unknown parameters or states to be combined with a first-principles model (26; 27; 28; 29; 30). The resulting hybrid model shows superior model accuracy compared to the first-principles model. But the field of neural networks has evolved from the use of a single hidden layer in an ANN to the use of multiple hidden layers in a DNN which requires exponentially less number of neurons than their shallow counterparts to approximate a specific function (31). Additionally, a hybrid modeling approach has never been applied to a full-scale bio-fermentation reactor.

Motivated by these limitations, we developed a DNN-based hybrid model approach for a full-

scale bio-fermentation process with a volume of over 100,000 gallons. Specifically, prior to building the hybrid model, we first improved the first-principles model by adding additional components and parameters to its equations based on process knowledge acquired from literature studies. This improved first-principles model was tested against multiple experimental datasets provided by an industry sponsor. Then, we identified critical parameters in the improved first-principles model which highly influence its outputs using local and global sensitivity analysis. Finally, these time-varying parameters were then estimated using a data-clustering approach, and approximated using a DNN which was then combined with the first-principles model to build a hybrid model (32). This hybrid model's performance was tested against multiple batches of process data provided by the industry sponsor, and compared against the accuracy of the first-principles model and the improved first-principles model.

Based on this developed hybrid model, the next step was building an observer to accurately track the states of the bio-fermentation process. An initial step in the design of the observer was assessing the performance of the traditional nonlinear Kalman filters, i.e., the Extended Kalman filter (EKF), and the Unscented Kalman filter (UKF) (33), to determine if they could be used to track the different states (34; 35). But these traditional methods were not able to accurately capture the nonlinearity of the states in this process model. Thus, in order to handle inter-sampling, i.e., assuming the available measurements are fairly accurate (little noise), another method was adopted. We developed an open-loop observer which utilized new sets of measurements, whenever they become available, to re-initialize the open-loop observer with the measured values (36). The developed observer showed superior estimation accuracy compared to the traditional Kalman filter and is computationally less expensive. The performance of the observer was then tested against other batches of process data.

As the final step of this work, an optimal control algorithm was developed in order utilizing the hybrid model to estimate the optimal operating conditions of the bio-fermenter in real-time. It is essential to maximize the product amount and minimize the Primary unit (Cost) for the profitability and productivity of the plant. It is also important to take practical constraints into account while

achieving these targets and computing the optimal operating conditions of the bio-fermenter to maintain an optimal substrate concentration throughout the process. And optimization problem was developed on GAMS which maximized the product and minimized the cost while considering the practical constraints. These optimal operating conditions are compared to the desired targets and the obtained state concentrations are plotted against the historical plant measurements.

The remainder of this thesis is organized as follows: Chapter 2 provides a brief background about DNNs, Levenberg-Marquardt algorithm, the methodology proposed by Bangi and Kwon (2020) to build a DNN-based hybrid model, a bio-fermentation process, and its first-principles model. It also describes the ideas used to improve the first-principles model using process knowledge. Chapter 3 presents the sensitivity analysis and clustering of parameters. It also presents building of the hybrid model for the bio-fermentation process using the improved first-principles model. Chapter 4 presents the observer design and the optimal control algorithm. Chapter 5 presents the concluding remarks.

2. Background

2.1 Deep neural networks

DNNs are neural networks with more than one hidden layer with each layer containing multiple neurons. The neurons in each layer are connected to all the neurons in the adjacent layers. Each connection between any two neurons carries a parameter called weight w, each layer has a bias b, and each neuron is associated with a function called activation function f.

Let $n^{k+1}(i)$ be the cumulative input received by unit i in layer $k+1$ which is given as

$$n^{k+1}(i) = \sum_{j=1}^{S_k} w^{k+1}(i,j)a^k(j) + b^{k+1}(i) \tag{2.1}$$

where S_k is the number of neurons in layer k, and $a^k(j)$ is the output of unit j in layer k which is given as

$$a^k(j) = f^k(n^k(j)), \qquad j = 1, ..., S_k \tag{2.2}$$

Assuming there are M layers in the network, the equations in matrix form can be represented as

$$A^k = F^k(W^k A^{k-1} + B^k), \qquad k = 0, 1, ..., M-1 \tag{2.3}$$

$$A^0 = u_q, \qquad q = 1, 2, ..., Q \tag{2.4}$$

where u_q is the input vector given to the neural network whose corresponding output is A_q^M obtained at the final layer M. The column vectors A^k, F^k, and B^k contain the outputs, the activation functions, and biases of all the neurons in layer k, respectively. The matrix W^k contains weights associated with the neurons in layers k and $k-1$, i.e., each row contains all the weights associated with a particular neuron of layer k, and the number of columns is equal to the number of neurons in layer $k-1$. The aim of the DNN is to learn the functional relationship between input and output

pairs $\{(u_1, y_1), (u_2, y_2), ...(u_Q, y_Q)\}$. The accuracy of the DNN is measured as follows:

$$V = \frac{1}{2} \sum_{q=1}^{Q} e_q^T e_q \tag{2.5}$$

$$e_q = y_q - A_q^M \tag{2.6}$$

where e_q is the error between the predicted output A_q^M and the actual output y_q when the q^{th} input (i.e., u_q) is fed to the DNN. The error matrix E can be defined as follows:

$$E = [e_1 \, e_2 \ldots e_Q]^T \tag{2.7}$$

2.2 Levenberg-Marquardt training

The Levenberg-Marquardt algorithm (37; 38) is a combination of two other parameter update algorithms, i.e., the Steepest Descent method and the Gauss-Newton algorithm. Consequently, it combines the stability characteristic of the steepest descent algorithm and the fast convergence of the Gauss-Newton algorithm. The update rule of the Levenberg-Marquardt algorithm is given by

$$w_{k+1} = w_k - (J_k^T J_k + \mu_{Train} I)^{-1} J_k E_k \tag{2.8}$$

$$b_{k+1} = b_k - (J_k^T J_k + \mu_{Train} I)^{-1} J_k E_k \tag{2.9}$$

where μ_{Train} is the combination coefficient used in training of the parameters, I is the identity matrix, and J is the Jacobian matrix which is defined as:

$$J = \begin{bmatrix} \frac{\partial e_1}{\partial w_1} & \frac{\partial e_1}{\partial b_1} & \frac{\partial e_1}{\partial w_2} & \cdots & \frac{\partial e_1}{\partial w_N} & \frac{\partial e_1}{\partial b_N} \\ \frac{\partial e_2}{\partial w_1} & \frac{\partial e_2}{\partial b_1} & \frac{\partial e_2}{\partial w_2} & \cdots & \frac{\partial e_2}{\partial w_N} & \frac{\partial e_2}{\partial b_N} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial e_Q}{\partial w_1} & \frac{\partial e_Q}{\partial b_1} & \frac{\partial e_Q}{\partial w_2} & \cdots & \frac{\partial e_Q}{\partial w_N} & \frac{\partial e_Q}{\partial b_N} \end{bmatrix} \tag{2.10}$$

where N is the total number of parameters involved in the learning of the DNN, and Q is the total number of data points used to train the DNN.

2.3 Hybrid model

Consider a dynamical system whose states are x, inputs are u, and outputs are y, and its dynamics is represented as:

$$y = h(x, u, p) \tag{2.11}$$

To build a hybrid model, a DNN will be trained to predict an unknown parameter p. This DNN is combined with the process model as shown in Fig. 2.1 to build a hybrid model (39). Before the training process can be initialized, certain parameters of the DNN are needed to be defined such as the number of layers, number of neurons in each layer, types of activation functions, and the initial values of weights and biases. Some of the commonly used activation functions are Sigmoid, Hyperbolic tangent, Rectified linear unit (ReLU), and Leaky rectified linear unit (Leaky ReLU). A point to remember is that if there is a lot of variation in the input-output variables, then it is better to normalize them as this helps speed up the DNN training process (40).

2.4 Training algorithm

An input-output training dataset with data points (u_1, y_1), (u_2, y_2), ..., (u_q, y_q), ..., (u_Q, y_Q) is used for training the hybrid model. The inputs are presented to the DNN and the first-principles model, and the outputs of the hybrid model $y^{'}$ are obtained. During this calculation, the output from the DNN, i.e., parameter values p, are obtained using Eqs. (2.1)-(2.4). These predictions are used as additional inputs to the first-principles model to calculate the outputs of the hybrid model $y^{'}$. During the training of the hybrid model, the squared prediction error of the outputs for all Q data points was minimized as follows:

$$\hat{V} = \frac{1}{2}\sum_{q=1}^{Q}(e_q)^T(e_q) \tag{2.12}$$

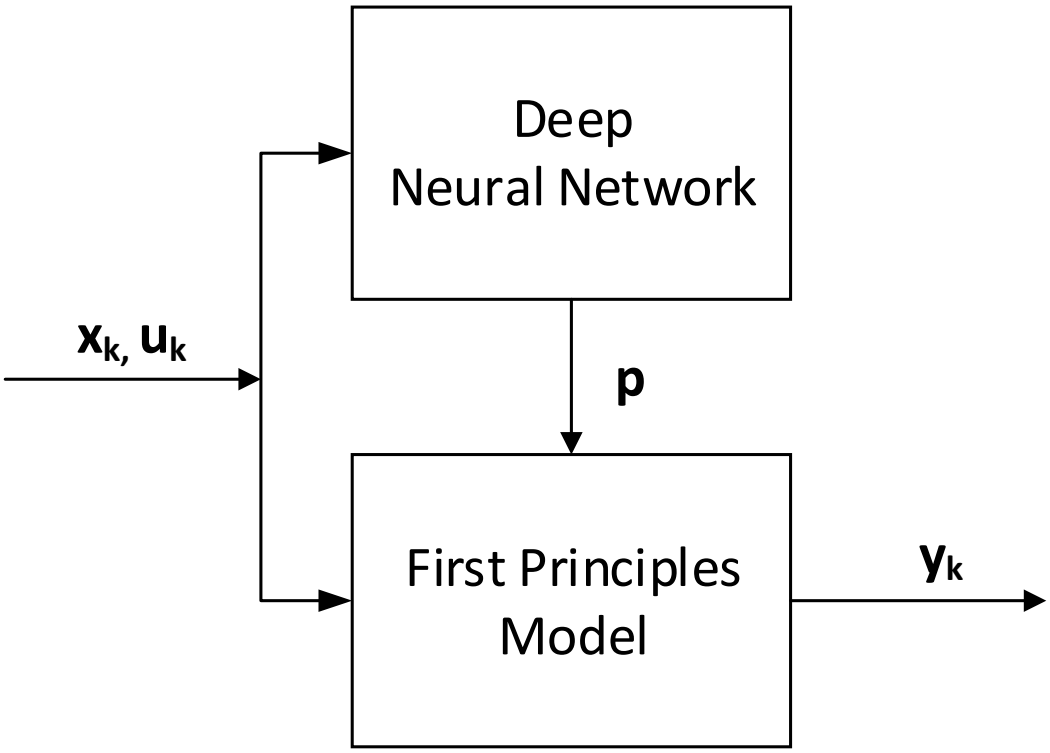

Figure 2.1: A schematic of the proposed hybrid model.

$$e_q = y_q - y_q^{'} \tag{2.13}$$

The parameters of the DNN are updated using Eqs. (2.8) and (4.2). But in order to calculate the Jacobian, the effect of the predicted parameter p_q on error e_q needs to be quantified indirectly as the DNN's output p_q is not directly related to e_q. To deal with this issue, we use a finite difference method to calculate the gradient of the hybrid model's output $y_q^{'}$ with respect to the DNN's output p_q, and is shown below:

$$\frac{\partial e_q}{\partial y_q^{'}} = -1 \tag{2.14}$$

$$\frac{\partial y_q^{'}}{\partial p_q} = \frac{y_{q+1}^{'} - 2y_q^{'} + y_{q-1}^{'}}{p_{q+1} - p_{q-1}} \tag{2.15}$$

$$\frac{\partial e_q}{\partial p_q} = \frac{\partial e_q}{\partial y_q^{'}} \frac{\partial y_q^{'}}{\partial p_q} = -\frac{y_{q+1}^{'} - 2y_q^{'} + y_{q-1}^{'}}{p_{q+1} - p_{q-1}} \tag{2.16}$$

The sensitivity of the error e_q to changes in the cumulative input of unit i in layer k is defined as:

$$\delta_q^k(i) = \frac{\partial e_q}{\partial n_q^k(i)} \tag{2.17}$$

Using Eq. (2.4), the above equation can be rewritten as:

$$\delta_q^k(i) = \frac{\partial e_q}{\partial n_q^k(i)} = \frac{\partial e_q}{\partial a_q^k(i)} \frac{\partial a_q^k(i)}{\partial n_q^k(i)} = \frac{\partial e_q}{\partial a_q^k(i)} \dot{f}^k(n_q^k(i)) \tag{2.18}$$

The above equation for the last layer M is as follows:

$$\delta_q^M = \frac{\partial e_q}{\partial A_q^M} \dot{F}^M(n_q^M) \tag{2.19}$$

But the output from the final layer is the predicted parameter p_q. Consequently,

$$\frac{\partial e_q}{\partial p_q} = \frac{\partial e_q}{\partial A_q^M} \tag{2.20}$$

Therefore, using Eqs. (2.16), (2.19), and (2.20), the δ_q^M value can be calculated. Now, the Jacobian matrix contains the sensitivities of the error e_q with respect to the parameters of the DNN, i.e., W^k and B^k. These sensitivities can be calculated for the parameters associated with the final layer M using δ_q^M and Eq. (2.1) as follows:

$$\frac{\partial e_q}{\partial W^M} = \delta_q^M A_q^{M-1} \tag{2.21}$$

$$\frac{\partial e_q}{\partial B^M} = \delta_q^M \tag{2.22}$$

For the parameters in the other layers, $k = 1, \ ..., \ M-1$, δ_q^k value can be calculated using the following recurrence relation:

$$\delta_q^k = \dot{F}^k(n_q^k) \ {W^{k+1}}^T \delta_q^{k+1} \tag{2.23}$$

Using δ_q^M value, the sensitivities of error e_q with respect to the parameters in the other layers can be calculated using Eqs. (2.21), (2.22) and (2.23). Once the calculation of the Jacobian matrix is completed, the parameters of the DNN can be updated using Eq. (2.8). The parameter μ_{Train} in the Levenberg-Marquardt training algorithm is not constant during the training process. It starts with an initial value but is multiplied with a factor β whenever a parameter update would result in the increase of $\hat{V}$ value. On the other hand, whenever a parameter update reduces the $\hat{V}$ value, then μ_{Train} is divided by β. The training algorithm is carried out until the value of $\hat{V}$ converges to a desired value. The flow diagram of the proposed hybrid modeling framework is presented with details in Fig. 2.2.

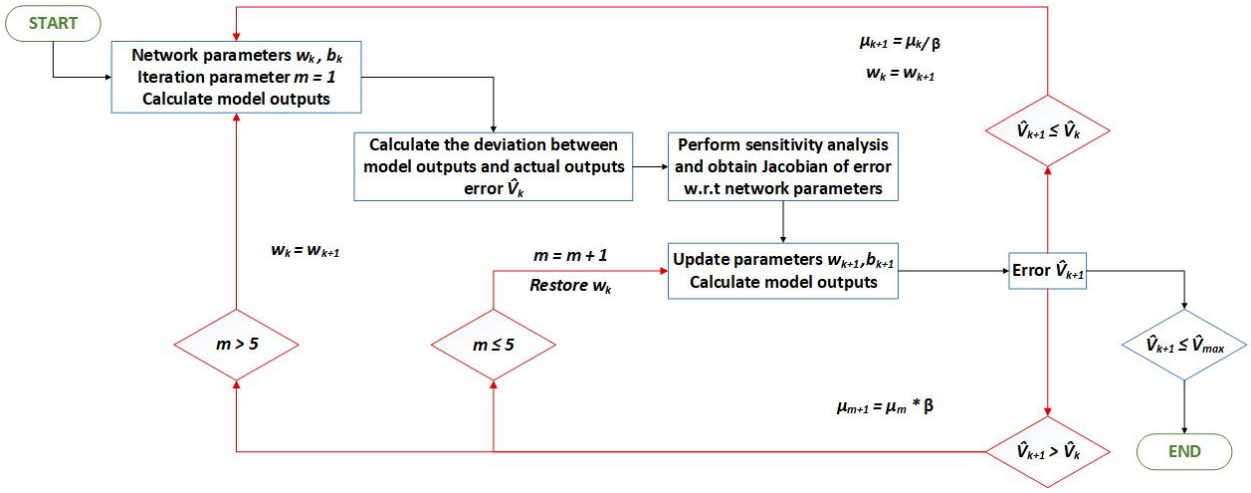

Figure 2.2: A block diagram for Levenberg-Marquardt based hybrid model training.

2.5 First-principles model of the bio-fermentation process

The first-principles model presented and discussed throughout this work is based on real process data provided by an industry sponsor, where the original first-principles model is similar to one used in (41). The growth rate model is as follows:

$$\mu = \mu_{S_1} + \mu_{S_2} + \mu_I \tag{2.24}$$

$$\mu_{S_1} = \frac{\mu_{max,S_1} \cdot \xi_{S_2} \cdot \xi_I \cdot S_1}{K_{S,S_1} + S_1 + a_{S_1,S_2} \cdot S_2 + a_{S_1,I} \cdot I} \tag{2.25}$$

$$\mu_{S_2} = \frac{\mu_{max,S_2} \cdot S_2}{K_{S,S_2} + S_2 + a_{S_2,S_1} \cdot S_1 + a_{S_2,I} \cdot I} \tag{2.26}$$

$$\mu_I = \frac{\mu_{max,I} \cdot I}{K_{S,I} + I + a_{I,S_1} \cdot S_1 + a_{I,S_2} \cdot S_2} \tag{2.27}$$

where, μ, μ_i, and $\mu_{max,i}$ refer to the overall growth rate, the growth rates associated with each component (i.e., Substrate 1 (S_1), Substrate 2 (S_2), Intermediate (I)), and the maximum specific growth rate of the micro-organisms associated with each component, respectively. $K_{S,i}$ and $a_{i,j}$ refer to the half-velocity constant associated with each component, and the inhibitory effect of component i on utilization of component j by the micro-organisms, respectively. It is important to note that $\xi_{S_2} \cdot \xi_I$ in Eq. (2.25) is incorporated with μ_{max,S_1} to account for any inhibitory effects Substrate 2 and Intermediate have on the growth rate associated with Substrate 1. The process has two operation modes, i.e., phase 1 and phase 2, and since the nutrient source is different in these two phases, the respective reactor models for each phase are different.

During phase 1, the reactor model can be described by the following equations:

$$\mu = \mu_{S_1} \tag{2.28}$$

$$\mu_{S_1} = \frac{\mu_{max,S_1} \cdot \xi_{S_2} \cdot \xi_{S_I} \cdot S_1}{K_{S,S_1} + S_1 + a_{s_1,s_2} \cdot S_2 + a_{s_1,I} \cdot I} \tag{2.29}$$

$$\frac{dB}{dt} = (\mu_{S_1} + \mu_{S_2} + \mu_I) \cdot B \tag{2.30}$$

$$\frac{dS_1}{dt} = -\frac{\mu_{S_1} \cdot B}{Y_{B/S_1}} \tag{2.31}$$

where B refers to Biomass, and Y_{B/S_1} refers to the yield coefficient of Biomass associated with Substrate 1.

During phase 2, the reactor model can be described by the following equations:

$$\frac{dB}{dt} = (\mu_{S_1} + \mu_{S_2} + \mu_I) \cdot B - mp_1 \cdot \frac{F_{in}}{V} \cdot B \tag{2.32}$$

$$\frac{dS_1}{dt} = -\frac{\mu_{S_1} \cdot B}{Y_{B/S_1}} - \frac{F_{in}}{B} \cdot S_1 \tag{2.33}$$

$$\frac{dS_2}{dt} = k_1 \cdot \mu_{S_1} \cdot B - \frac{\mu_{S_2} \cdot B}{Y_{B/S_2}} - \frac{F_{in}}{V} \cdot (S_2 - S_{2_{initial}}) \tag{2.34}$$

$$\frac{dI}{dt} = (k_2 \cdot \mu_{S_1} + k_3 \cdot \mu_{S_2}) \cdot B - \frac{\mu_I \cdot B}{Y_{B/I}} - \frac{F_{in}}{V} \cdot I \tag{2.35}$$

$$\frac{dP}{dt} = (\alpha_1 \cdot \mu_{S_1} + \alpha_2 \cdot \mu_{S_2} + \alpha_3 \cdot \mu_I) \cdot B + \beta \cdot B - \frac{F_{in}}{V} \cdot P \tag{2.36}$$

$$\frac{dV}{dt} = F_{in} \tag{2.37}$$

where P is the Product concentration, V is the reactor volume, α_i is the coefficient linked to the growth rate responsible for increase in Product, k_i refers to the coefficient linked to the growth rate responsible for the increase in Substrate 2 and Intermediate, β is the coefficient linked to the non-growth associated term responsible for the increase in Product, and F_{in} refers to the feed flow rate of Substrate 2. The parameters Y_{B/S_1}, Y_{B/S_2}, and $Y_{B/I}$ refer to the yield coefficients of Biomass associated with each component. It should be noted that the coefficients linked to the growth rate, k_i, incorporate temperature dependence through Arrhenius equation as follows:

$$k_i = c_i \cdot e^{\frac{-E_{a_i}}{RT}} \tag{2.38}$$

where c_i is the pre-exponential factor, E_{a_i} is the activation energy, R is the universal gas constant, and T refers to the temperature.

It is important to note that this is the original model that was provided by the industry sponsor, who also introduced a manipulated parameter (mp_1) to the second term of Eq. (2.32) associated with Biomass in order to obtain a better fit. The following section will provide a summary of modifications that were carried out on the original first-principles model to improve its prediction capability. It is also important to note that all results presented in this work are normalized, at the request of the industry sponsor.

2.6 Improving the first-principles model

The previous sections provides a background of deep neural networks (DNNs), the Levenberg Marquadt training algorithm, and the hybrid model that will be used in the bio-fermentation process. A description of the original first-principles model for this process, specifically the growth rate model, and the reactor models for phase 1 and phase 2, was also presented.

Experimental data was utilized in order to carry out estimation of the growth rate parameters and the overall growth rate for the original first-principles model. These parameters were estimated using fmincon optimization in MATLAB by minimizing the normalized sum of squared error (SSE) of all the outputs. A summary of the estimated growth rate parameters for the original first-principles model is provided in Table 2.1. A comparison of the estimated overall growth rate to the experimental overall growth rate is illustrated in Chapter 3, where the model estimate obtained using Eqs. (2.24)-(2.28) is observed to be accurate, despite some variability in the experimental values. It is important to note that the growth rate parameters were assumed to be constant throughout the bio-fermentation process.

Table 2.1: Estimated growth rate parameters for the original first-principles model.

Parameter	Value	Unit
μ_{max,S_1}	0.775	hr^{-1}
K_{S,S_1}	2.30×10^2	g Substrate 1 L^{-1}
μ_{max,S_2}	0.544	hr^{-1}
K_{S,S_2}	4.16×10^2	g Substrate 2 L^{-1}
$\mu_{max,I}$	0.545	hr^{-1}
$K_{S,I}$	2.82×10^2	g Intermediate L^{-1}
a_{s_1,s_2}	1.94×10^2	–
$a_{s_1,I}$	1.94×10^2	–
a_{s_2,s_1}	1.93×10^2	–
$a_{s_2,I}$	1.95×10^2	–
a_{I,s_1}	2.08×10^2	–
a_{I,s_2}	1.95×10^2	–

The yield coefficient for Substrate 1 is assumed to differ in each phase. Assuming the yield coefficient is the only unknown during phase 1, the reactor model for phase 1 is solved in order to obtain the yield coefficient, i.e., $Y_{B/S_1} = 0.749$. A comparison of the estimated states to experimental data for phase 1 is provided in Fig. 2.3, where the model estimates obtained are shown to be accurate.

Similarly, the reactor model for phase 2 was solved in order to obtain estimates of concentrations of Biomass, substrates, etc., as illustrated in Fig. 2.4. Although the model values for Biomass, Substrate 1, Product, Intermediate, and Volume are reasonable, a downward trend can be observed for Substrate 2. Substrate 2 is a raw material and related to operating costs, and thus, accurate estimation of this state is essential. A summary of the kinetic parameters used in the original first-principles model for phase 2 is provided in Table 2.2. It should also be noted that the industry sponsor desires accurate estimation of Product and Biomass as well.

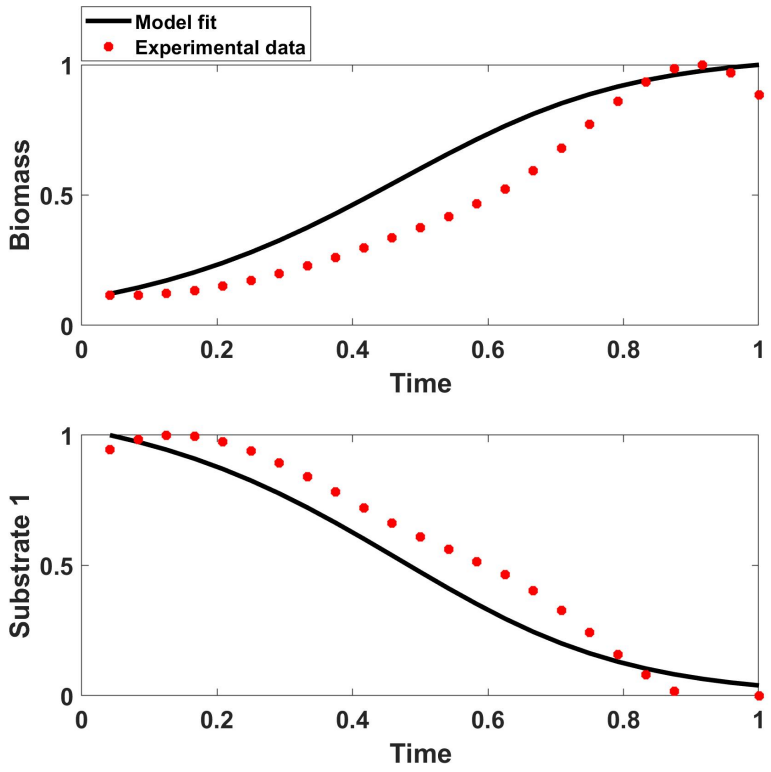

Figure 2.3: A comparison of the original first-principles model and training data during phase 1.

A second set of experimental data was used for validation purposes employing the model parameters obtained using the original experimental data, and these results are presented in terms of root mean squared error (RMSE) values in the next chapter and are illustrated in Fig. 3.1. It can be observed that the model estimates for Biomass, Substrate 1, Product, and Volume are reasonable. However, a downward trend can once again be observed for Substrate 2. In addition, a large discrepancy is observed between the model and experimental values for Intermediate. As the estimation results for the original first-principles model are not ideal for certain states, an improvement in the original first-principles model was first pursued through the incorporation of two additional components. These components may account for discrepancies encountered when utilizing the original first-principles model, and these modifications will be discussed next. These modifications include the incorporation of two components X_1 and X_2. Component X_1 is a manipulated input for which continuous values are available, and it behaves similar to catalyst when

Table 2.2: Estimated phase 2 parameters for the original first-principles model.

Parameter	Value	Unit
Y_{B/S_1} (phase 2)	0.184	g Cell/g Substrate 1
Y_{B/S_2}	0.212	g Cell/g Substrate 2
$Y_{B/I}$	0.250	g Cell/g Intermediate
c_1	0.250	g Substrate 2/g Cell
c_2	0.250	g Intermediate/g Cell
c_3	0.045	g Intermediate/g Cell
α_1	0.100	g Product/g Substrate 1
α_2	0.085	g Product/g Substrate 2
α_3	0.013	g Product/g Intermediate
β	2.14×10^{-5}	g Product/g Cell.hr
Ea_1	0.100	J/mol
Ea_2	0.100	J/mol
Ea_3	9.27×10^2	J/mol
$mp1$	0.005	–

added during phase 2 of the process. Component X_2 is an essential chemical which ensures optimal conditions for micro-organisms. Measurements of component X_2 are also available, and the incorporation of both components in the original first-principles model is highly desirable.

2.6.1 Incorporation of component X_1

Through information obtained from historical operation data, the addition of component X_1 in phase 2 was observed to increase the production of the micro-organisms, consequently resulting in an increase in the consumption of Substrate 2, and thus, an increase in the production of Product. This information was utilized to update Eqs. (2.34) and (2.36) through the addition of empirical terms as follows:

$$\frac{dS_2}{dt} = k_1 \cdot \mu_{S_1} \cdot B - \frac{\mu_{S_2} \cdot B}{Y_{B/S_2}} - \frac{F_{in}}{V} \cdot (S_2 - S_{2_{initial}}) - p_1 \cdot X_1 \tag{2.39}$$

$$\frac{dP}{dt} = (\alpha_1 \cdot \mu_{S_1} + \alpha_2 \cdot \mu_{S_2} + \alpha_3 \cdot \mu_I) \cdot B + \beta \cdot B - \frac{F_{in}}{V} \cdot P + p_2 \cdot X_1 \tag{2.40}$$

where p_1 and p_2 are empirical coefficients that allow X_1 to be incorporated as consumption and production terms in the equations for Substrate 2 and Product, respectively.

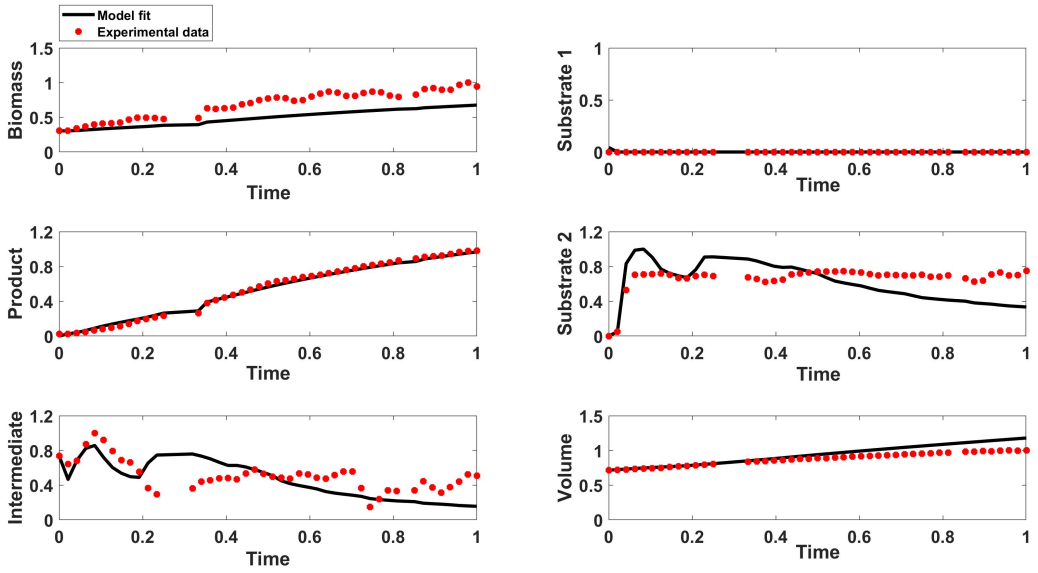

Figure 2.4: A comparison of the original first-principles model and training data during phase 2.

It is important to note that since X_2 functions in a way similar to catalyst, it was initially incorporated in the growth rate coefficient related to μ_{S_2} in Eq. (2.26). Unfortunately, satisfactory results were not obtained through this approach, and the empirical terms shown in Eqs. (2.39) and (2.40) had to be introduced instead.

2.6.2 Incorporation of component X_2

Through information obtained from historical operation, it is understood that measurements from an online sensor are available for X_2. Therefore, it is desirable to incorporate component X_2 as well. X_2 is expected to play a role similar to the role played by oxygen in bio-fermentation processes, and was hence incorporated as follows (42):

$$\mu = (\mu_{S_1} + \mu_{S_2} + \mu_I) \cdot \left(\frac{X_2}{K_{X_2} + X_2} \right) \tag{2.41}$$

$$\frac{dX_2}{dt} = k_L a \cdot (X_{2_{max}} - X_2) - q_{X_2} \cdot B \tag{2.42}$$

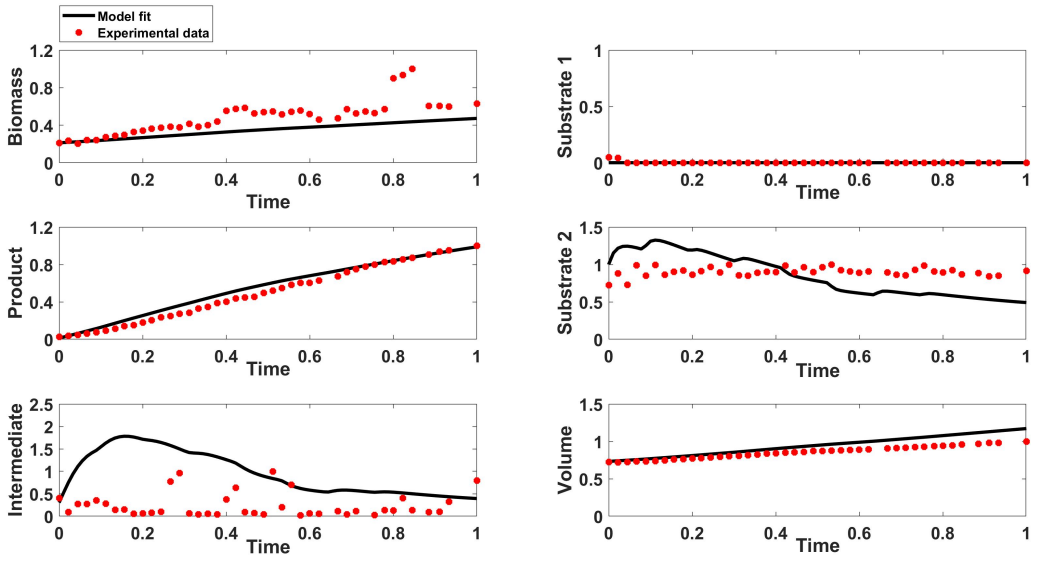

Figure 2.5: A comparison of the original first-principles model and validation data during phase 2.

where $X_{2_{max}}$ and K_{X_2} are the maximum value of X_2 during a particular batch run, and the half-velocity constant associated with X_2, respectively. $k_L a$ and q_{X_2} are the mass transfer coefficient, and uptake rate of X_2, respectively. It should be noted that online measurements for X_2 are available throughout phase 1 and phase 2 of the bio-fermentation process. An additional objective of the industry sponsor was to incorporate X_2 in the original first-principles model and predict it like the other states, so that its estimates can be tracked in real-time.

Experimental data was utilized in order to estimate the growth rate parameters and the overall growth rate for the revised first-principles model. A comparison of the experimental overall growth rate and the overall growth rate available from the original and revised first-principles model is provided in the next chapter, where estimates from both the original and revised model are shown to be accurate. A summary of the estimated growth rate parameters for the revised model is provided in Table 2.3.

The revised first-principles model can now be solved for phase 1, and a comparison of the model fit to experimental data is provided in Fig. 2.6. The corresponding model parameter values are provided in Table 2.4. Model estimates are shown to be accurate for all states. Two additional

Table 2.3: Estimated growth rate parameters for the revised first-principles model.

Parameter	Value	Unit
μ_{max,S_1}	0.808	hr^{-1}
K_{S,S_1}	2.44×10^2	g Substrate 1 L^{-1}
μ_{max,S_2}	0.377	hr^{-1}
K_{S,S_2}	1.84×10^2	g Substrate 2 L^{-1}
$\mu_{max,I}$	0.954	hr^{-1}
$K_{S,I}$	63.7	g Intermediate L^{-1}
a_{s_1,s_2}	5.08×10^2	–
$a_{s_1,I}$	9.99×10^2	–
a_{s_2,s_1}	4.91×10^2	–
$a_{s_2,I}$	1.01×10^2	–
a_{I,s_1}	4.95×10^2	–
a_{I,s_2}	5.13×10^2	–
K_{X_2}	6.49×10^{-4}	hr^{-1}

Table 2.4: Estimated phase 1 parameters for the revised first-principles model.

Parameter	Value	Unit
Y_{B/S_1} (phase 1)	0.903	g Cell/g Substrate 1
k_La (phase 1)	0.117	g Substrate 1 hr^{-1}
q_{X_2} (phase 1)	0.223	–

experimental datasets were utilized in order to validate the revised first-principles model for phase 1 and compare it with first-principles model using RMSE values. It is seen in Table 2.5 that RMSE values for the model estimates for all the states, especially Substrate 1 and Biomass, are reasonably good. This demonstrates the significantly improved accuracy of the revised first-principles model for phase 1.

Similarly, the revised first-principles model can now be solved for phase 2, and a comparison of the model and the experimental data is provided in Fig. 2.7. The corresponding model parameter values are provided in Table 2.6. It should be noted that Substrate 1 is 0 in phase 2 (see Fig. 2.4 and Fig. 2.5), as it is consumed completely in phase 1, and is therefore not included in the plots for phase 2 for the remainder of this work. While the model estimates for Biomass, Product, X_2, and Volume are fairly accurate, discrepancies can be noted in the model estimates for Substrate 2 and Intermediate. It should be noted that while there may be inaccuracies in the estimation of

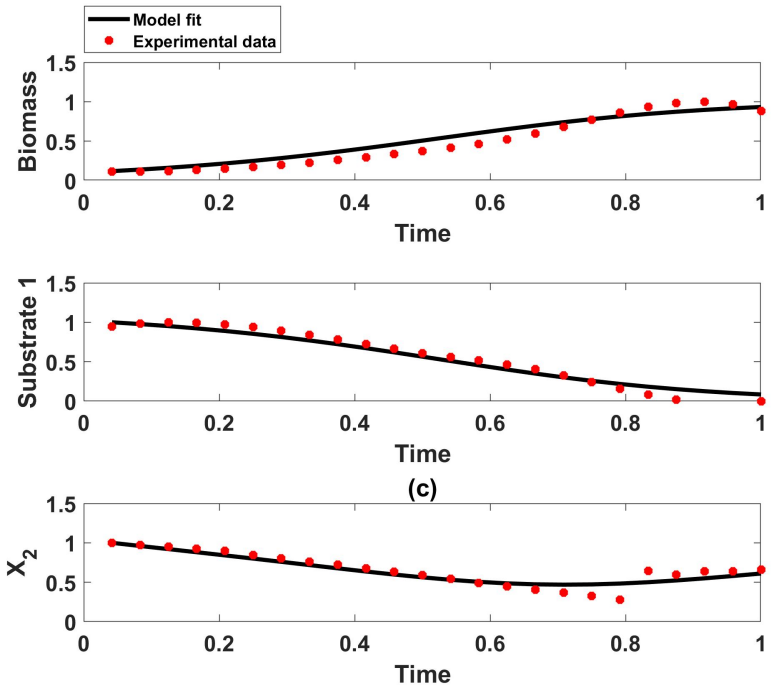

Figure 2.6: A comparison of the revised first-principles model and training data during phase 1.

Table 2.5: RMSE values for both models using training data and two validation datasets during phase 1

Dataset	Model	Substrate 1	Biomass	X_2
Training data	First-principles model	0.1299	0.1273	-
	Revised first-principles model	0.0959	0.0870	0.0754
Validation dataset 1	First-principles model	0.1235	0.3553	-
	Revised first-principles model	0.0927	0.1060	0.2155
Validation dataset 2	First-principles model	0.2260	0.4163	-
	Revised first-principles model	0.0978	0.144	0.1345

Intermediate, it may be attributed to the complexity of the revised first-principles model due to incorporation of added components.

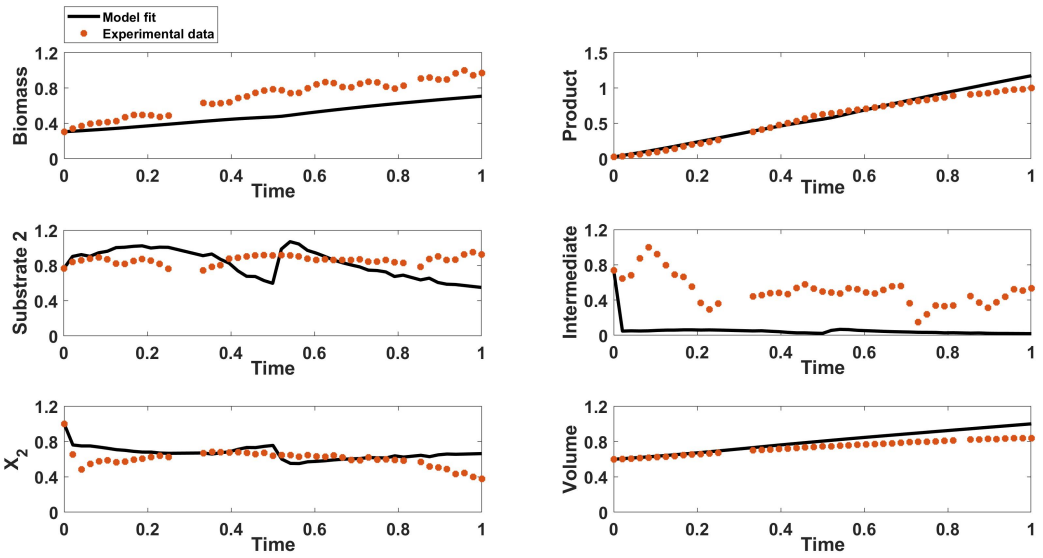

Figure 2.7: A comparison of the revised first-principles model and training data during phase 2.

Table 2.6: Estimated phase 2 parameters for the revised first-principles model.

Parameter	Value	Unit
Y_{B/S_1} (phase 2)	0.321	g Cell/g Substrate 1
Y_{B/S_2}	0.459	g Cell/g Substrate 2
$Y_{B/I}$	0.012	g Cell/g Intermediate
c_1	0.108	g Substrate 2/g Cell
c_2	0.108	g Intermediate/g Cell
c_3	0.036	g Intermediate/g Cell
α_1	0.173	g Product/g Substrate 1
α_2	0.040	g Product/g Substrate 2
α_3	0.156	g Product/g Intermediate
β	1.34×10^{-4}	g Product/g Cell·hr
Ea_1	24.8	J/mol
Ea_2	24.8	J/mol
Ea_3	0.498	J/mol
p_1	9.99×10^{-5}	hr^{-1}
p_2	1.80×10^{-7}	hr^{-1}
k_La	5.13	–
q_{X_2}	16.6	–
$mp1$	1	–

Additional experimental data was utilized in order to validate the revised first-principles model

for phase 2 (Fig. 2.8) and quantified using RMSE values in chapter 3. These results further highlight the inability of the revised first-principles model to accurately estimate Substrate 2, particularly at the start of phase 2, even though Intermediate is estimated reasonably well in this case. The results also show that X_2 is predicted reasonably well but it is important to further improve the accuracy. Since the estimation of Substrate 2 is crucial, a hybrid model can improve the estimation, and a sensitive analysis, clustering, and a hybrid modeling approach is presented in the following chapter to guide its development.

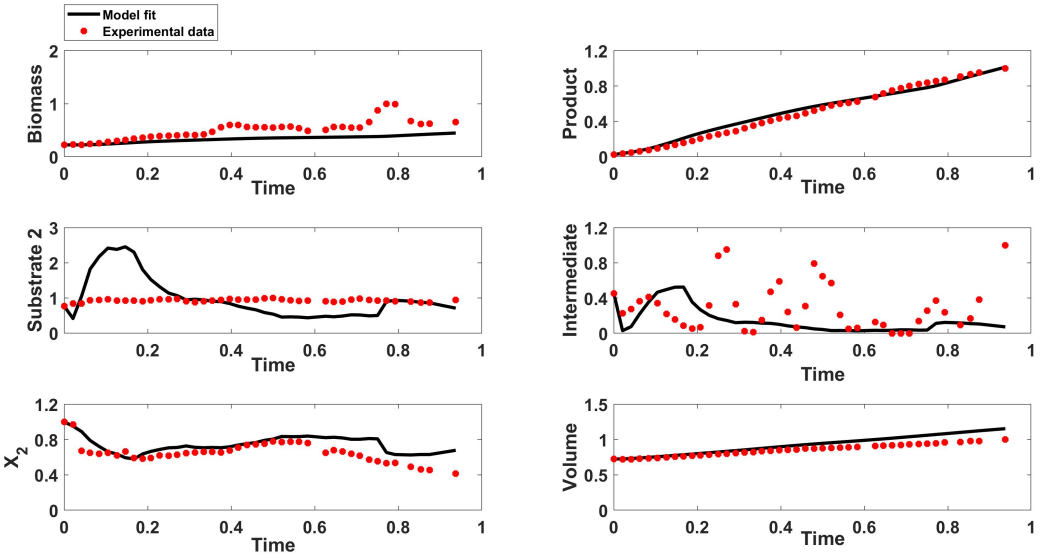

Figure 2.8: A comparison of the revised first-principles model and validation data during phase 2.

3. Sensitivity analysis, and a hybrid modeling approach

To develop a hybrid model that utilizes all available information (i.e., first-principles), sensitivity analysis needs to be carried out to identify highly sensitive model parameters.

3.1 Sensitivity analysis

Before modifying the model further by considering new measurements or using data-driven approaches to improve the existing model, it is essential to perform a sensitivity analysis to understand which parameters greatly influence the outputs. In this section, a local and global sensitivity analysis of the model is presented.

A local sensitivity analysis around the nominal values of the model parameters is initially carried out to understand how the model parameters and initial conditions influence the different outputs, i.e., Substrate 1, Substrate 2, Biomass, Product, Intermediate, and X_2.

A sensitivity matrix is first derived in order to come up with the parameter set which affects the outputs (43). The sensitivity matrix shows the dependencies of the outputs with respect to the parameters and initial conditions, as shown below:

$$S = \begin{bmatrix} \partial y_1(t_1)/\partial\theta_1 & \dots & \partial y_1(t_1)/\partial\theta_{n_\theta} \\ \vdots & \ddots & \vdots \\ \partial y_1(t_{n_t})/\partial\theta_1 & \dots & \partial y_1(t_{n_t})/\partial\theta_{n_\theta} \\ \vdots & \ddots & \vdots \\ \partial y_{n_y}(t_1)/\partial\theta_1 & \dots & \partial y_{n_y}(t_1)/\partial\theta_{n_\theta} \\ \vdots & \ddots & \vdots \\ \partial y_{n_y}(t_{n_t})/\partial\theta_1 & \dots & \partial y_{n_y}(t_{n_t})/\partial\theta_{n_\theta} \end{bmatrix} \tag{3.1}$$

Here, yϵR^{n_y} represents the output states, and $\theta\epsilon R^{n_\theta}$ represents the parameters and initial condition whose sensitivity analysis is carried out (44). These sensitivity matrix values are typically normalized by multiplying with the nominal values of parameters and by dividing through the nominal

values of the outputs to ensure that different units for the parameters/outputs do not affect the sensitivity analysis results.

To capture the effect that these parameters have on the outputs, a criterion called Fisher information matrix is calculated in the form of sensitivity matrix as follows:

$$FIM = S^T \Sigma S \tag{3.2}$$

where Σ is an identity matrix. A specific criterion is required to evaluate the information contained in the Fisher information matrix, and for this purpose, the D-optimality criterion (ϕ_D) is used. It minimizes the logarithm of the determinant of the inverse of the Fisher information matrix. Using inverse determinant property, we get:

$$\phi_D^* = max\ \phi_D\ (FIM) = max\ log\ det\ (FIM) \tag{3.3}$$

For the purpose of this sensitivity analysis, one parameter is evaluated at a time, and ϕ_D is computed to show the effect it has on an output. A higher ϕ_D value implies that the concerned model parameter has a higher influence on the given output. As the process states in the first-principles model are measurable, these states are the model outputs in the bio-fermentation process, i.e., y = x where x is the state. It is important to note that the study is carried out for the outputs of the reactor model for phase 2, since it is the most important phase of the bio-fermentation process as Substrate 2 is present in this phase and majority of Product is formed in this phase. Substrate 2 is tied to the operating cost of the process and is the main energy source for the micro-organisms. As the supply of Substrate 1 is limited, it is primarily used for the initial growth of Biomass, and only a relatively small percentage of product is formed in phase 1 when compared to phase 2. To calculate the sensitivity of states with respect to the parameters, the following equation is solved:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial x}{\partial \theta_i} = \frac{\partial f}{\partial x^T}\frac{\partial x}{\partial \theta_i} + \frac{\partial f}{\partial \theta_i} \tag{3.4}$$

When θ represents an initial condition of the state in Eq. (3.4), the second term ($\frac{\partial x}{\partial \theta_i}$) is 1, and

Table 3.1: Local sensitivity analysis: a list of sensitive parameters with D-optimality criterion (ϕ_D) values for (a) when output states in the model for phase 2 are equally important, and (b) when Substrate 2 and Product are 5 times more weighted than the other states.

Equal weight to outputs	ϕ_D	Substrate 2/Product with 5 times weight	ϕ_D
$S_{2_{initial}}$	22.7	V_0	68.6
Y_{B/S_2}	22.6	$S_{2_{initial}}$	68.01
V_0	22.3	Y_{B/S_2}	67.8
B_0	9.02	α_2	27.4
μ_{max,S_2}	8.66	μ_{max,S_2}	25.9
K_{S,S_2}	8.24	K_{S,S_2}	24.6
α_2	5.47	B_0	6.08
man_para	0.858	β	2.08
β	0.416	G_0	0.018
G_0	0.018	P_0	-5.77

for all other parameters, this term is 0 (43). For the case of this bio-fermentation process, the dimension of the overall sensitivity matrix is $[195, 6, 38]$ where 195 denotes the number of time instants in the process, 6 denotes the number of states (i.e. Substrate 1, Biomass, Product, Substrate 2, Intermediate, added component X_2), and 38 is the number of parameters including the initial conditions. Overall, in this sensitivity analysis study, we consider the effect of these 38 parameters on the output states. The parameters consist of six initial conditions corresponding to the six states, and 32 parameters from the growth rate and phase 2 reactor model. The nominal values of these parameters are the values listed in Tables 2.1 and 2.2.

Since the local sensitivity analysis was carried out initially, all the parameters were considered to be at the nominal values. The result from this study is shown in Table 3.1. Two cases are examined, the first case is where all the output states are assumed to be equally important, and the second case is where Substrate 2 and Product are considered 5 times more weighted than the rest, as they are the primary states of interest and need to be predicted accurately. Parameters are listed according to decreasing order of ϕ_D for both the cases. The importance of the parameter is determined by how high it appears in the table.

For the first case where output states are assumed to be equally weighted, it is seen that the yield coefficient associated with Substrate 2 and the initial concentration of Substrate 2 being fed into

the reactor are particularly important, along with the initial conditions of the state. Additionally, μ_{max,S_2}, K_{S,S_2}, and α_2 are also considerably important. The initial conditions of the output and the initial concentration of pure Substrate 2 flowing into the fermenter, $S_{2_{initial}}$, cannot be estimated using optimization. This is because, their value is subject to the real-time operation of the bio-fermentation plant, varying with different batches. Thus, the main focus is on the parameters present in the growth rate and phase 2 that can be better estimated in order to attain a superior output prediction compared to the revised first-principles model.

From the local sensitivity analysis, it can be concluded that regardless of the weight to the states, the initial concentration of Substrate 2 in the feed, the yield coefficient with respect to Substrate 2, and the initial value of Volume are important parameters. The initial conditions of most of the states are important. The half-velocity constant K_{S,S_2}, which is associated with Substrate 2, is the only half-velocity constant that is important. The non-associated growth term β is an important parameter affecting Product. The maximum specific growth rate associated with Substrate 2, μ_{max,S_2}, and the growth rate coefficient associated with Substrate 2 responsible for the increase in Product, α_2, also play important roles in affecting the outputs.

As the parameters generally vary a lot depending on different batches, operating conditions, and changes in measurements, it is important to see how sensitive the outputs are, based on a wide range of parameter values, i.e., through a global sensitivity analysis. From global sensitivity analysis, we can identify the parameters and initial conditions that are the most important and significantly affect the outputs, particularly, Substrate 2 and Product concentrations. To decide the range of parameter values upon which the global sensitivity analysis model would be developed, 5 experimental datasets were run with a slight change in certain conditions, e.g., initial parameter guess, parameter bounds, etc. Based on the estimated parameter values obtained from these runs, an overall range was decided for each parameter and initial condition, leaving a 20-50% margin of error to account for the maximum range of values possible. A wider range is generally preferred because different batches can estimate different parameter values, and it is helpful to see how the outputs might react to parameter values that are far from their nominal value.

For the global sensitivity analysis, we use $lhsdesign$ (Latin Hypercube Sampling) in MATLAB to construct a matrix of random values between 0 to 1 for each parameter. 100 different cases of parameter sets are considered for this analysis and are averaged at the end. The overall combined global sensitivity analysis of the model is done using the formulation below:

$$\begin{aligned}\phi_D(model) = \phi_D(Substrate\ 1) + \phi_D(Biomass) + \phi_D(Intermediate) \\ +5 \cdot [\phi_D(Substrate\ 2) + \phi_D(Product)]\end{aligned} \tag{3.5}$$

Table 3.2 lists the parameters along with their lower and upper bounds (LB and UB) that were used for global sensitivity analysis. The range of each parameter considered is shown below:

$$Range\ of\ each\ parameter = [LB \quad LB + (UB - LB) \cdot lhsdesign] \tag{3.6}$$

Table 3.2: A list of parameters with lower and upper bounds for global sensitivity analysis.

No.	Parameter	LB	UB
1	Y_{B/S_1}	0.01	0.25
2	α_1	0	0.25
3	β	10^{-5}	10^{-3}
4	c_1	0	0.25
5	Y_{B/S_2}	0.01	0.25
6	c_2	0	0.25
7	$Y_{B/I}$	0.01	0.25
8	α_2	0	0.25
9	α_3	0	0.25
10	c_3	0	0.25
11	S_{1_0}	0	2
12	B_0	50	70
13	P_0	0.1	1.2
14	S_{2_0}	200	3000
15	I_0	0	0.2
16	V_0	4×10^4	1.2×10^5
17	$S_{2_{initial}}$	500	1200
18	μ_{max,S_1}	0.1	1.2
19	μ_{max,S_2}	0.4	1.2
20	$\mu_{max,I}$	0.4	1.2
21	Ea_1	0	25
22	Ea_2	0	1000
23	Ea_3	0	1200
24	man_para	0.001	1.5
25	K_{S,S_1}	0	500
26	K_{S,S_2}	100	1000
27	$K_{S,I}$	100	4000
28	a_{s_1,s_2}	100	8000
29	$a_{s_1,I}$	0	4000
30	a_{s_2,s_1}	0	4000
31	$a_{s_2,I}$	100	4000
32	a_{I,s_1}	100	4000
33	a_{I,s_2}	100	4000
34	p_1	1×10^{-6}	1×10^{-4}
35	p_2	1×10^{-8}	1×10^{-6}
36	K_{X_2}	0	0.01
37	k_La	0	50
38	q_{X_2}	0	50

The effect of all these 38 parameters on the outputs is studied, individually and for the model as a whole. The results of global sensitivity analysis for individual outputs are summarized in Table 3.3, where the importance of the parameter is determined by how high it appears in the table. Most of the parameters in Table 3.3 have a positive ϕ_D value and parameters not included in this table are negative, implying that they do not affect the outputs significantly.

It should be noted that the parameters are much more significant for the global sensitivity analy-

Table 3.3: Global sensitivity analysis: a list of sensitive parameters for each output present in the revised first-principles model for phase 2.

No.	Substrate 1	Biomass	Product	Substrate 2	Intermediate	Added component X_2
Parameters						
1	Y_{B/S_1}	Y_{B/S_2}	α_2	$a_{S_2,I}$	$a_{S_2,I}$	V_0
2	G_0	E_0	Y_{B/S_2}	E_0	μ_{max,S_2}	k_La
3	μ_{max,S_1}	$S_{2_{initial}}$	E_0	μ_{max,S_2}	E_0	$S_{2_{initial}}$
4	–	$a_{S_2,I}$	$S_{2_{initial}}$	Y_{B/S_2}	Y_{B/S_2}	q_{X_2}
5	–	man_para	β	a_{S_2,S_1}	a_{S_2,S_1}	Y_{B/S_2}
6	–	B_0	$a_{S_2,I}$	K_{S,S_2}	$a_{S_2,I}$	p_1
7	–	μ_{max,S_2}	μ_{max,S_2}	$S_{2_{initial}}$	K_{S,S_2}	Y_{B/S_1}
8	–	a_{S_2,S_1}	V_0	V_0	$S_{2_{initial}}$	μ_{max,S_2}
9	–	V_0	c_3	man_para	c_3	K_{S,S_2}
10	–	c_3	a_{S_2,S_1}	a_{I,S_2}	$\mu_{max,I}$	B_0
11	–	K_{S,S_2}	P_0	B_0	$Y_{B/I}$	E_0
12	–	$Y_{B/I}$	K_{S,S_2}	c_3	man_para	$Y_{B/I}$

sis, compared to the local sensitivity analysis where the initial conditions were of high significance to the outputs. It is seen that the yield coefficient associated with Substrate 2, initial concentration of Substrate 2 when it is fed into the fermenter, inhibition parameters $a_{S_2,I}$ (effect of Substrate 2 on utilization of Intermediate by micro-organisms), and maximum specific growth rate of the micro-organisms associated with Substrate 2 are the most important parameters. Additionally, β (non-growth associated term responsible for the increase in Product) and α_2 (coefficient associated with Substrate 2 responsible for the increase in Product) are sensitive to Product. K_{S,S_2} (Half velocity associated with Substrate 2) is sensitive to Substrate 2. Only three parameters are seen to affect Substrate 1 in phase 2, and the rest of the parameters have a negative or zero ϕ_D value.

In combined global sensitivity analysis of the developed revised first-principles model, the effect of all outputs is considered together. Substrate 2 and Product outputs are given five times more weight than Substrate 1, Biomass, and Intermediate. As mentioned earlier, the reason for that is the need for accurate prediction of these two states. A summary of the results of the combined global sensitivity analysis is shown in Table 3.4. From Table 3.4, the following conclusions can be made:

- Regardless of the weight to the states and global/local analysis, it should be noted that the

Table 3.4: Global sensitivity analysis: a list of sensitive parameters with D-optimality criterion (ϕ_D) values when Substrate 2 and Product are 5 times more weighted than the other states.

Substrate 2/Product with 5 times weight	ϕ_D
E_0	65.0
$a_{S_2,I}$	59.9
Y_{B/S_2}	55.4
μ_{max,S_2}	38.6
α_2	37.34
$S_{2_{initial}}$	29.4
β	16.9
a_{S_2,S_1}	6.80
K_{S,S_2}	-7.14
P_0	-8.74

initial concentration of Substrate 2 in the feed, $S_{2_{initial}}$, the yield coefficient with respect to Substrate 2, Y_{B/S_2}, coefficient associated with Substrate 2 responsible for the increase in Product, α_2, and maximum specific growth rate associated with Substrate 2, μ_{max,S_2}, are important parameters.

- The initial condition of all states is important for local analysis, but not as important for global analysis.

- The half-velocity constant associated with Substrate 2, K_{S,S_2}, appears to be the only half-velocity constant of significant importance.

- Inhibition parameters are very important and sensitive to the model according to the global sensitivity analysis, especially the ones associated with Substrate 2.

- The non-associated growth term, β, is an important parameter affecting Product.

The results of the sensitivity analysis can be utilized in order to examine variation in sensitive model parameters, through parameter clustering as presented in the following section.

3.1.1 Improving the revised first-principles model through clustering

The sensitivity analysis identified μ_{max,S_2}, $\mu_{max,I}$, and K_{S,S_2} as the sensitive model parameters for the growth rate, and Y_{B/S_2}, c_3, α_2, and k_La as the sensitive model parameters for phase 2. Some of the other sensitive parameters like $a_{S_2,I}$ and other inhibition parameters are sensitive to the outputs but they do not vary with time. The identified model parameters can now be utilized in a parameter clustering approach to observe their variation through the course of the bio-fermentation process, and to determine which parameters might benefit from a hybrid model approach. This approach is beneficial if there is no first-principles model to define potential time-varying parameters. In these cases, DNNs are used to develop a relation between frequently available online measurements and estimated parameters. This allows the model accuracy to be improved by utilizing time-varying parameters rather than a single estimate for the given model parameter.

To the knowledge of the authors, there were no first-principles models for any of the sensitive model parameters that were identified, and thus, a clustering approach was pursued to determine if there were large variations in the sensitive model parameters. In this approach, sensitive parameters are estimated separately in different clusters of time. This approach provides different estimates for the sensitive model parameter in each cluster, thus enabling time-varying parameters to be obtained.

Experimental datasets provided by the industry sponsor included the values at 50 different time instants for each state. These were used to create 5 clusters, each comprised of 10 values. Insensitive growth parameters were fixed to the values provided in Table 2.1. This is done since limited experimental data is available, and re-estimating all parameters may lead to over-fitting.

A comparison of the simulation results of all the developed models, i.e., first-principles model, revised first-principles model, and clustered model with experimental data for the growth rate, is provided in Fig. 3.1. Here, improved estimation of the growth rate using the parameter clustering approach can be observed on the normalized time scale. But still the clustered model is unable to accurately track the time-varying nature of the growth rate characteristics. The parameters estimated by the clustered approach for the growth rate are presented in Table 3.5. Similarly, the

parameters estimated by the clustered approach for the reactor model for phase 2 are presented in Table 3.6. These results demonstrate that the sensitive model parameters are time-varying, particularly the growth rate coefficients associated with Substrate 2 and Intermediate, μ_{max,S_2} and $\mu_{max,I}$, and yield coefficient associated with Substrate 2, $Y_{B/S_2,1}$. These parameters may benefit from developing a hybrid model, which will be explored in the following section.

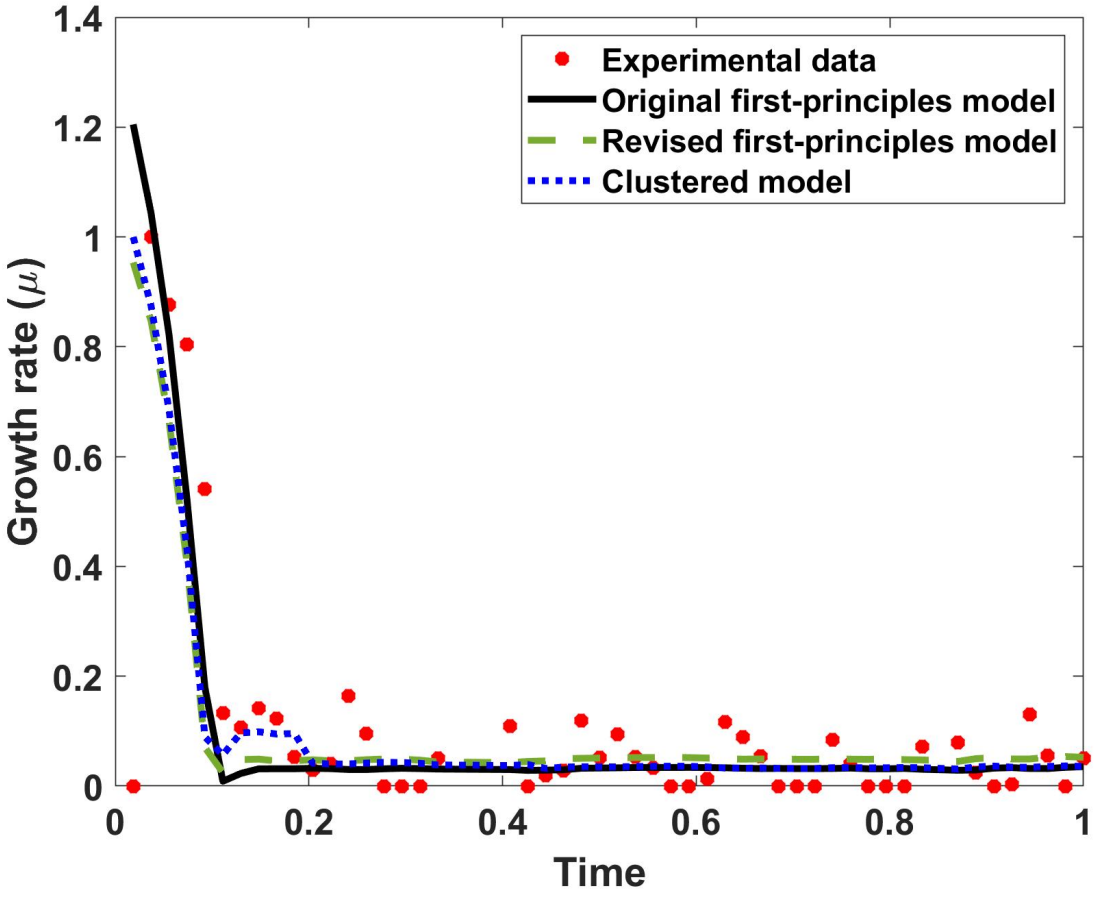

Figure 3.1: A comparison of growth rate parameter estimation using the first-principles model, revised first-principles model, and clustered model.

Table 3.5: Clustered growth rate parameters.

Growth rate parameter	Time period (normalized)	Value	Unit
$\mu_{max,S_2,1}$	0.0-0.2	0.512	hr^{-1}
$\mu_{max,S_2,2}$	0.2-0.4	0.202	hr^{-1}
$\mu_{max,S_2,3}$	0.4-0.6	0.133	hr^{-1}
$\mu_{max,S_2,4}$	0.6-0.8	0.124	hr^{-1}
$\mu_{max,S_2,5}$	0.8-1.0	0.129	hr^{-1}
$K_{S,S_2,1}$	0.0-0.2	68.9	g Substrate 2 L^{-1}
$K_{S,S_2,2}$	0.2-0.4	2.03×10^2	g Substrate 2 L^{-1}
$K_{S,S_2,3}$	0.4-0.6	2.37×10^2	g Substrate 2 L^{-1}
$K_{S,S_2,4}$	0.6-0.8	2.85×10^2	g Substrate 2 L^{-1}
$K_{S,S_2,5}$	0.8-1.0	2.79×10^2	g Substrate 2 L^{-1}
$\mu_{max,I,1}$	0.0-0.2	0.972	hr^{-1}
$\mu_{max,I,2}$	0.2-0.4	0.772	hr^{-1}
$\mu_{max,I,3}$	0.4-0.6	0.687	hr^{-1}
$\mu_{max,I,4}$	0.6-0.8	0.555	hr^{-1}
$\mu_{max,I,5}$	0.8-1.0	0.612	hr^{-1}

Table 3.6: Clustered phase 2 parameters.

Phase 2 parameter	Time period (normalized)	Value	Unit
$Y_{B/S_2,1}$	0.0-0.2	0.855	g Cell/g Substrate 2
$Y_{B/S_2,2}$	0.2-0.4	0.155	g Cell/g Substrate 2
$Y_{B/S_2,3}$	0.4-0.6	0.135	g Cell/g Substrate 2
$Y_{B/S_2,4}$	0.6-0.8	0.115	g Cell/g Substrate 2
$Y_{B/S_2,5}$	0.8-1.0	0.156	g Cell/g Substrate 2
$c_{3,1}$	0.0-0.2	0.30	g Intermediate/g Cell
$c_{3,2}$	0.2-0.4	1.00	g Intermediate/g Cell
$c_{3,3}$	0.4-0.6	1.50	g Intermediate/g Cell
$c_{3,4}$	0.6-0.8	1.50	g Intermediate/g Cell
$c_{3,5}$	0.8-1.0	1.50	g Intermediate/g Cell
$\alpha_{2,1}$	0.0-0.2	0.011	g Product/g Substrate 2
$\alpha_{2,2}$	0.2-0.4	0.056	g Product/g Substrate 2
$\alpha_{2,3}$	0.4-0.6	0.051	g Product/g Substrate 2
$\alpha_{2,4}$	0.6-0.8	0.035	g Product/g Substrate 2
$\alpha_{2,5}$	0.8-1.0	0.010	g Product/g Substrate 2
$k_L a_1$	0.0-0.2	13.4	hr^{-1}
$k_L a_2$	0.2-0.4	3.36	hr^{-1}
$k_L a_3$	0.4-0.6	3.36	hr^{-1}
$k_L a_4$	0.6-0.8	2.36	hr^{-1}
$k_L a_5$	0.8-1.0	2.06	hr^{-1}

3.2 Development of the hybrid model

3.2.1 Improving the revised first-principles model through hybrid modeling

In the previous section, sensitive model parameters were identified and estimated in a clustered manner, where each of the five values estimated for the parameters was used to improve model prediction. The parameters mentioned in Table 3.5 and Table 3.6 show that there was significant variation in their values with time, but parameters such as μ_{max,S_2}, $\mu_{max,I}$, and $Y_{B/S_2,1}$ change more frequently, and nonlinearly in time unlike k_La, and $c_{3,1}$. Thus, to get a more accurate representation of these parameters and capture their complete time-varying nature over all the time instants, a hybrid modeling approach was adopted.

As described in Chapter 2, a hybrid model is one that utilizes a data-driven model along with a first-principles model (45; 46). A DNN is trained to estimate μ_{max,S_2}, $\mu_{max,I}$, and $Y_{B/S_2,1}$, which are to be utilized in the improved first-principles model. The inputs to the DNN are the concentrations of Substrate 2, Biomass, Intermediate, Product, Volume, and X_2. As this approach is primarily for phase 2 model, Substrate 1 is not considered since it is experimentally known to be negligible during this phase as it gets consumed almost completely in phase 1. Even the model fit shows a negligible concentration of Substrate 1 as seen in Fig. 2.4 and Fig. 2.5. The first layer is the input, and the last layer is the output. The nodes are connected using weights, and each node has a bias. Rectified Linear (ReLu) activation function is used to calculate the output of each node. The DNN used in this work consists of 3 hidden layers with 5 nodes each and 3 outputs, which are the parameters mentioned above. These parameters are then used in the first-principles model, and the output concentrations are calculated.

Now, as shown in Fig. 2.1, x_k are the states of the improved first-principles model mentioned earlier, i.e., Substrate 2, Biomass, Product, Intermediate, X_2, and Volume, and u_k are the manipulated inputs to the process which will also be used as an input to the hybrid model such as temperature, added component X_1, alkali flow rate, and Substrate 2 flow rate. It is important to note that alkali flow rate is a critical input to the fermenter as it is added to keep the pH in check

as it can neutralize Intermediate and added chemical X_1.

The DNN is initially pre-trained using the MATLAB deep learning toolbox, and the clustering parameter values. These parameters are used in the first-principles model for phase 2 wherein the other parameters are constant. The output concentrations from the hybrid model are represented as x_{k+1}, which will be used as input to the model in the next time step. Y_k is the plant measurement of these output concentrations. Once we have the output from the hybrid model, the error is calculated using Eq. (2.13). Based on the error for all outputs, the SSE is calculated using Eq. (2.12). As mentioned in Section 2.2, Jacobian matrix is then calculated which is used to update the weights and biases. Levenberg Marquardt algorithm is used to update the DNN parameters. These updated weights and biases are then used in the next iteration, and the hybrid model gives a new set of outputs. Once again, the error is calculated, and SSE is computed. This process is repeated until the error is less than a tolerance value. It is important to note that unlike the clustering method which had 5 estimated values for each parameter, the DNN has 50 parameter values corresponding to 50 output measurements, thus estimating time-varying parameters in a much more accurate manner (47).

The results for the hybrid model using the training data are illustrated in Fig. 3.2. It can be seen that all the states are predicted well and the model fit is more accurate than the revised first-principles model's prediction, especially for Product and Substrate 2, which are the primary states of interest in this work. Estimation of states using validation data is shown in Fig. 3.3, and it shows that all the states except Intermediate are predicted fairly accurately. There is an order of magnitude difference between the measurements of Intermediate from different batches due to uncertainty from the yeast cells and the significant effects X_1 and alkali flow rate have on it. Hence, its prediction for validation batch is not as accurate as the other states. The main concern with Intermediate concentration is regarding identification of abnormality in bio-fermentation process, and there are other means of tracking Intermediate that the industry sponsor uses. The results also show that the prediction accuracy of component X_2 is high thus the developed model can be used to successfully track X_2 along with the other states.

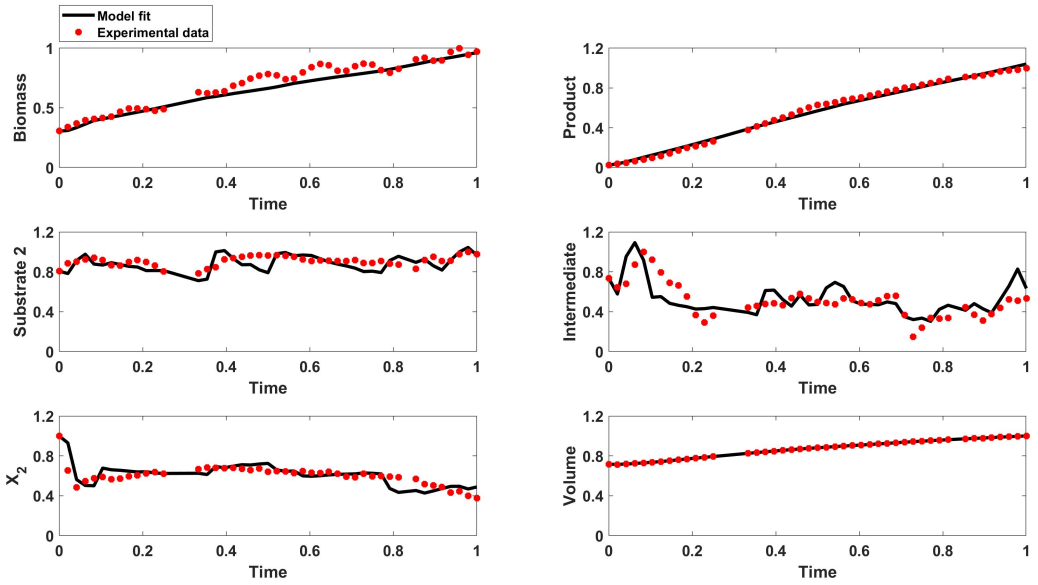

Figure 3.2: A comparison of the hybrid model and training data during phase 2.

The parameters estimated by the hybrid model were used to further validate 2 additional batches, and both showed accurate predictions of Substrate 2 and product, as shown in Fig. 3.4, and Fig. 3.5. Due to difficulty in taking offline measurements, only Substrate 2 and Product concentrations are measured during normal operation of the bio-fermentation plant. The results show reasonable prediction for both these states using the two batches. For further improving the prediction accuracy, it is crucial to have more experimental data so that the neural network can be trained even more precisely as larger the sample size of data, the better the parameter estimates. For the objective of this work, all these validation plots show that the states, especially Substrate 2 and Product, are predicted reasonably well.

3.2.2 Error analysis

In order to quantify and compare the performance of the three models, i.e., original first-principles model, revised first-principles model, and hybrid model, we utilize the relative error (RE) formulation as defined below:

$$RE_k = 1 - \frac{x_{k+1}}{Y_k} \tag{3.7}$$

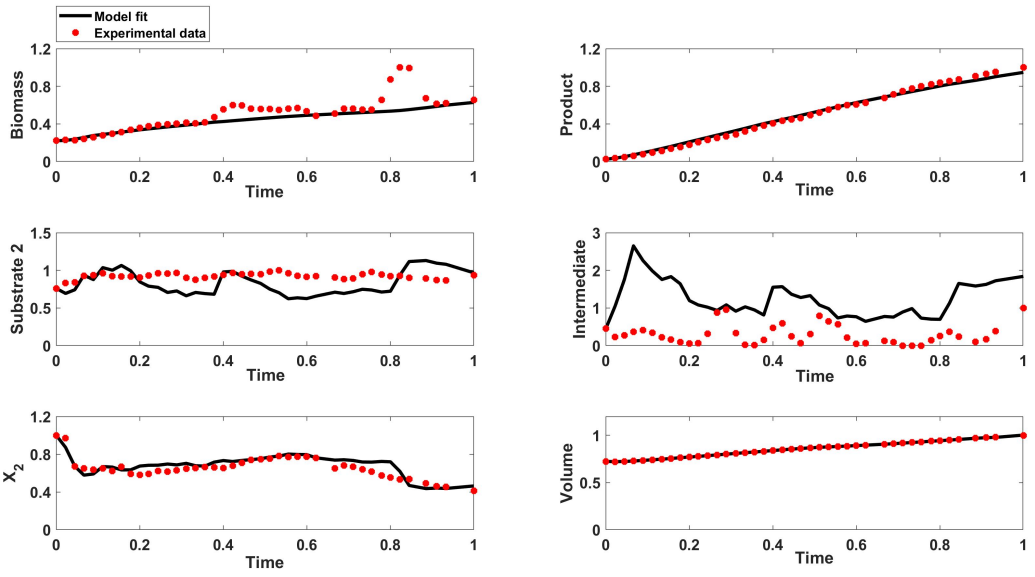

Figure 3.3: A comparison of the hybrid model and validation data during phase 2.

In Eq. (3.7), x_{k+1} are the predicted state concentrations using the models and Y_k are the plant measurements, as described in detail in the previous section. First, performance of the three models is compared using the training batch that was used to train the DNN in the hybrid model. The predicted outputs from these three models were utilized to calculate the RE value as defined in Eq. (3.7) at every time step. The RE plots for the training dataset are plotted and compared in Fig. 3.6. The key observation here is that the hybrid model outperforms the first-principles model and the revised first-principles model. This can be attributed to the fact that the hybrid model includes a trained DNN which accurately predicts the sensitive parameter values as well as the dependencies among themselves, and this results in better prediction of outputs.

Next, the performance of the three models is compared using the validation batch. The predicted outputs were utilized to calculate the RE value as defined in Eq. (3.7). Once again, the hybrid model performs much better than the other two models, except in the case of the prediction of Intermediate, where the performance of the hybrid model is comparable to the first-principles model.

Additionally, the error is numerically quantified using the root mean squared error (RMSE)

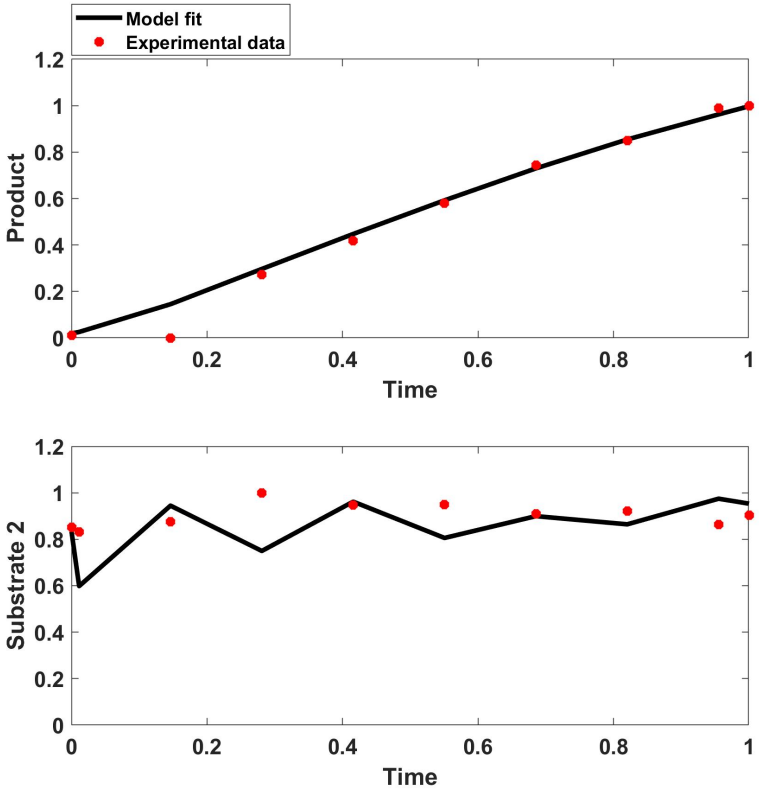

Figure 3.4: A comparison of the hybrid model and additional validation dataset 1, during phase 2.

formulation as defined below:

$$RMSE = \sqrt{\frac{\sum_{k=0}^{N-1} (Y_k - x_{k+1})^2}{N}} \tag{3.8}$$

RMSE values were calculated by comparing the predictions from all three models against the training and validation batches, and are summarized in Table 3.7 and Table 3.8, respectively. From these tables, it can be observed that the hybrid model outperforms the first-principles model and the revised first-principles model in the prediction of all the states except for the Intermediate. Moreover, the RMSE values for Product and Substrate 2 for the two additional validation batches using the hybrid model were found to be low: 0.0625 (Product) and 0.1279 (Substrate 2) for the first, and 0.0448 (Product) and 0.1240 (Substrate 2) for the second. In conclusion, the hybrid model shows superior performance as it is equipped with a DNN that predicts time-sensitive parameters

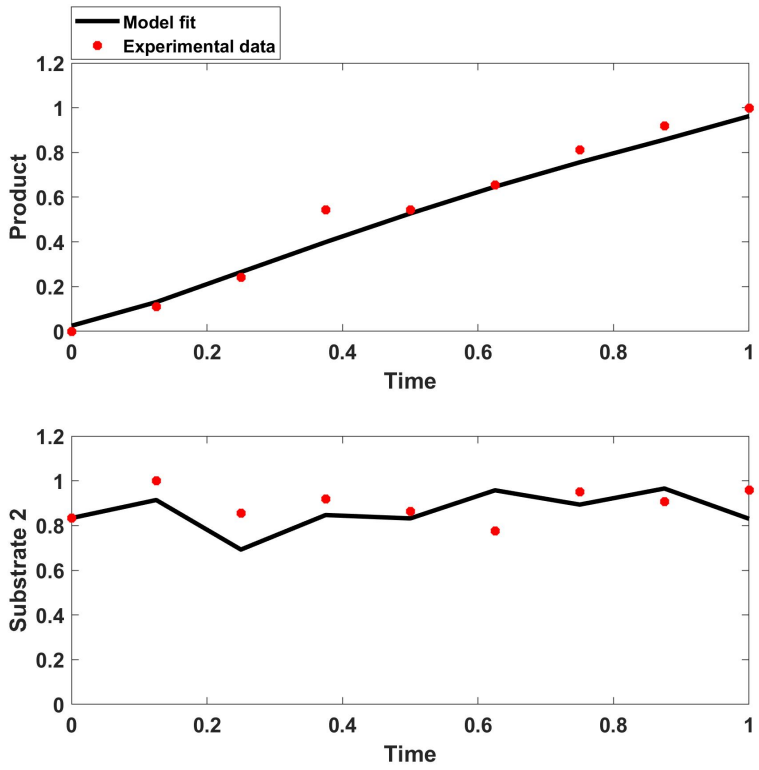

Figure 3.5: A comparison of the hybrid model and additional validation dataset 2, during phase 2.

accurately.

Table 3.7: RMSE values for all three models using training data

	Biomass	Product	Substrate 2	Intermediate	X_2	Volume
First-principles model	0.2162	0.0260	0.2886	0.2046	-	0.0792
Revised first-principles model	0.2218	0.0672	0.1932	0.4851	0.1158	0.0993
Hybrid model	0.0590	0.0278	0.0707	0.1368	0.0719	5.877e-04

Table 3.8: RMSE values for all three models using validation data

	Biomass	Product	Substrate 2	Intermediate	X_2	Volume
First-principles model	0.1972	0.0480	0.2968	2.536	-	0.0862
Revised first-principles model	0.2189	0.0463	0.6030	0.3460	0.1170	0.0776
Hybrid model	0.1274	0.0278	0.1862	1.079	0.0639	9.792e-04

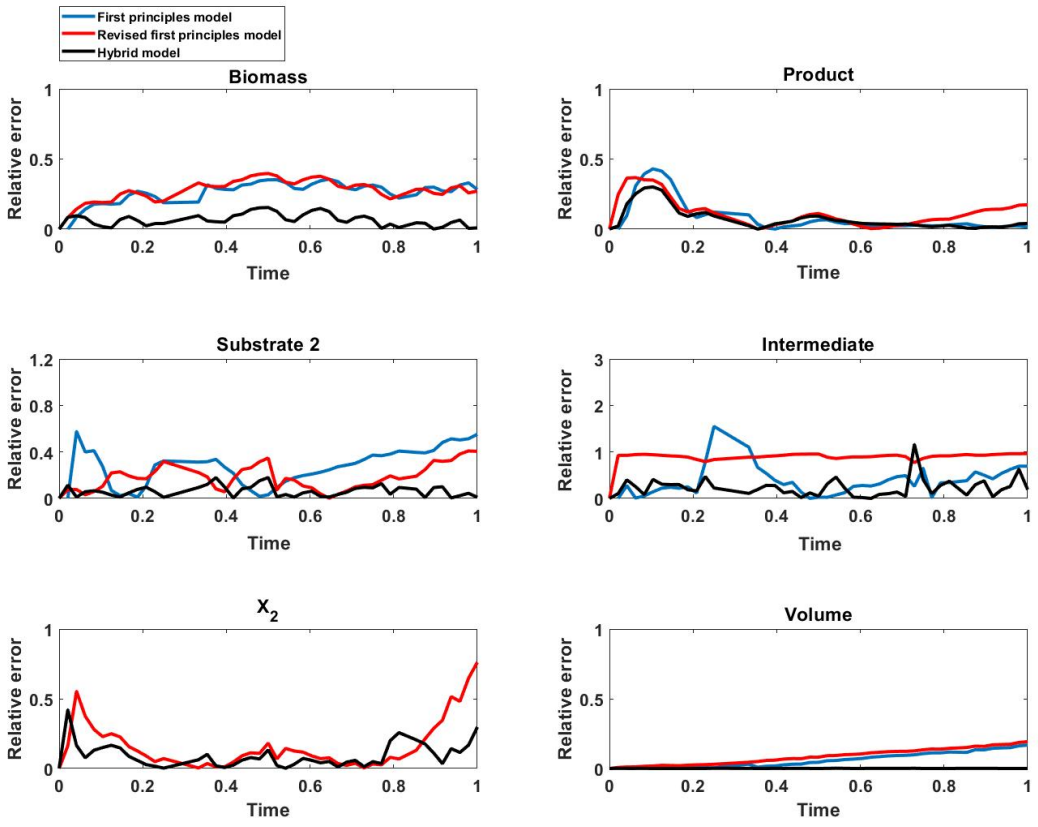

Figure 3.6: Relative errors between the models (i.e, the first-principles model, revised first-principles model, and hybrid model) and the training data obtained from the industry sponsor.

4. Observer Design and Optimal Control Algorithm

The previous chapter details the development of hybrid model after identifying the highly sensitive model parameters using sensitivity analysis and clustering. The next step in order to further towards running the model online is developing an observer which can track the internal states of the system. And after that, we develop an optimal control algorithm using an optimization problem which can be used in a model predictive framework.

4.1 Observer Design

The main goal of an observer is to accurately estimate internal states of the process between the sampled data (Inter-sample measurements). An initial step in the design of the observer was to assess the performance of the traditional nonlinear Kalman filters in order to determine if they could be used to estimate the states of the bio-fermentation process. The assumption here is that three states have measurements available: product, Substrate 2, and X_2. The goal is to determine if all the other states can be tracked using these available measurements. For better estimation, the measurements were interpolated and Forward-Euler method was used in the simulation. But the results showed that both EKF and UKF were unsuccessful in tracking biomass, and Intermediate.

Thus, in order to handle inter-sampling, an open-loop observer with re-initialization was used. The available measurements were assumed to be fairly accurate and with little noise present. This method utilized a new set of state measurements, whenever they become available, to re-initialize the open-loop observer with the measured values. This approach is computationally less demanding, and given the scarcity in measurements, it is a reasonable method to implement. The observer is designed for both phases and the states are re-initialized whenever an new measurement is available. For phase 1, Substrate 1, and X_2 were re-initialized while in phase 2, Substrate 2, Product, and X_2 were re-initialized.

It should be noted that the observer designed is multi-rate i.e all the states measurements do not need to be available at the same instant. Thus, as online measurements of X_2 and Volume are

available at a much higher frequency, they can be utilized as input to the observer.

A schematic illustration of the observer model is shown in Fig. 4.1. The observer model utilizes continuous input from the online sensors for volume, X_1, X_2, temperature, and Substrate 2 flow rate. The observer model also utilizes Substrate 2 and Product lab measurements whenever available in order to re-initialize the open-loop observer. States that are output from the model are Substrate 2, Product, Biomass, Intermediate, X_2, and volume.

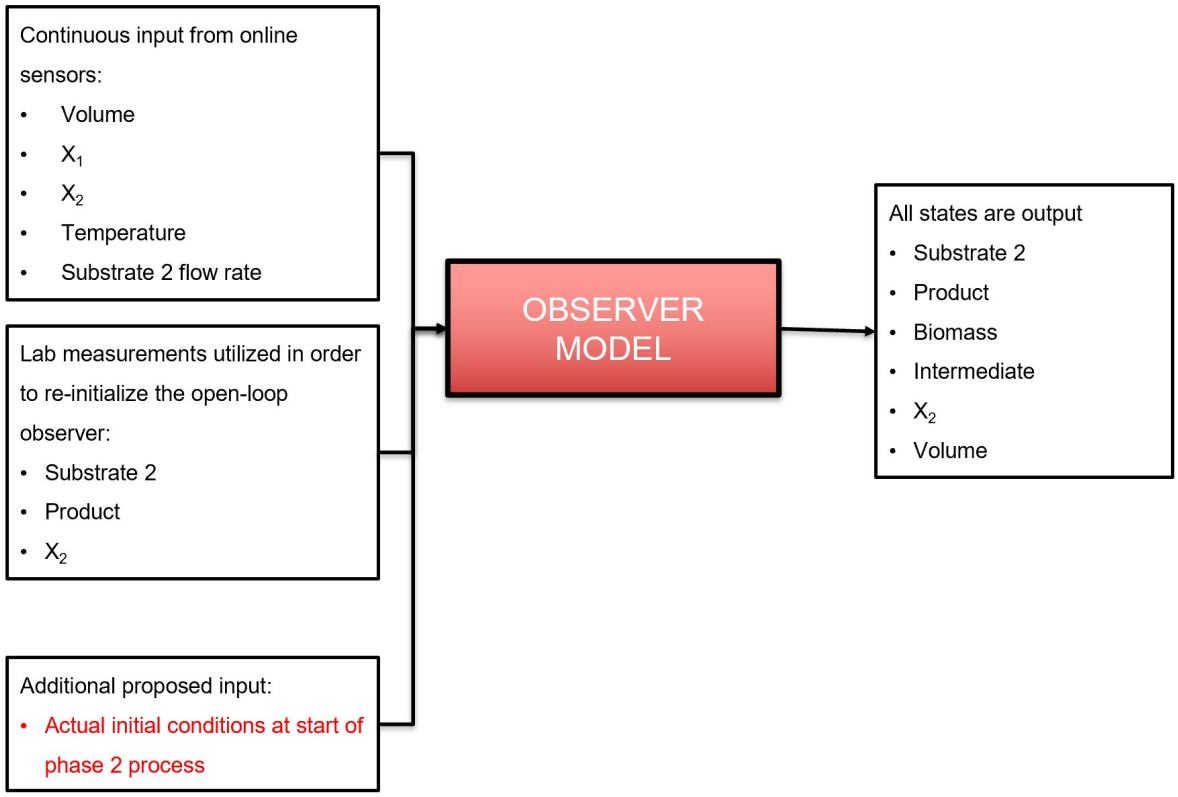

Figure 4.1: A schematic illustration of observer design.

The observer is implemented for data from training batch, and the plot for the time series evolution of all states is illustrated in Fig. 4.2 and Fig. 4.3. It should be noted that Substrate 2, Product, Volume, and X_2 states are only re-initialized every time when the plant measurements are available. The results show reasonable performance of the observer for all states for both phase

shown in Fig. 4.1, and for the overall process (phase 1 and phase 2 combined) shown in Fig. 4.2.

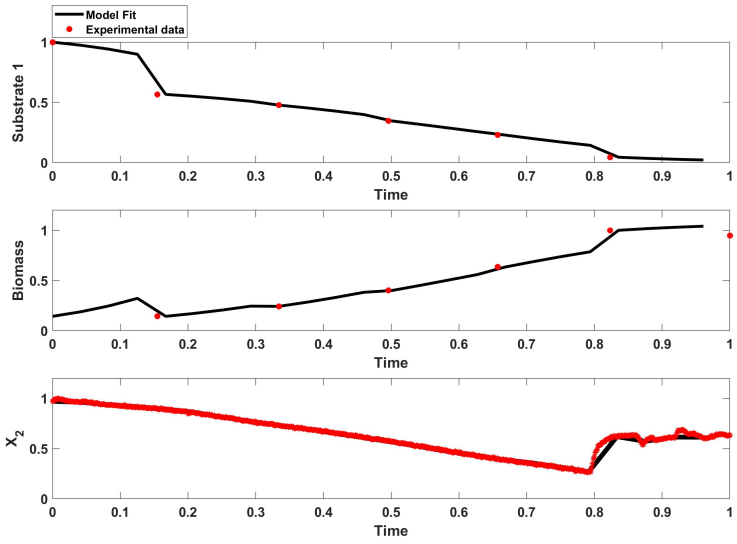

Figure 4.2: A comparison of the hybrid model-based observer fit and training data during phase 1.

Another validation dataset was used to test the developed hybrid model-based observer design. The plots for the time series evolution of all states is illustrated in Fig. 4.4 (Phase 1) and Fig. 4.5 (Combined phase 1 and 2). It is seen that all the three states in phase 1 are tracked accurately. In this case, for illustration purposes, even Intermediate state is re-initialized and we can see in Fig. 4.4 that all the six states are tracked well in between the samples thus showing the prediction accuracy of the developed observer.

4.2 Optimal control algorithm

The primary aim of any industrial process is increased productivity and reduced cost. In order to achieve that aim, many resources are used to optimize the process such that it runs at optimal operating conditions and maximises the profitability. The same is true for a bio-fermentation process. The motivation behind developing an optimization problem is to utilize the developed hybrid model to estimate the optimal operating conditions of the bio-fermenter in real-time. It is essential to maximize the product amount and minimize the Primary unit (Cost) for the profitability and productivity of the plant. It is also important to take practical constraints into account while achieving

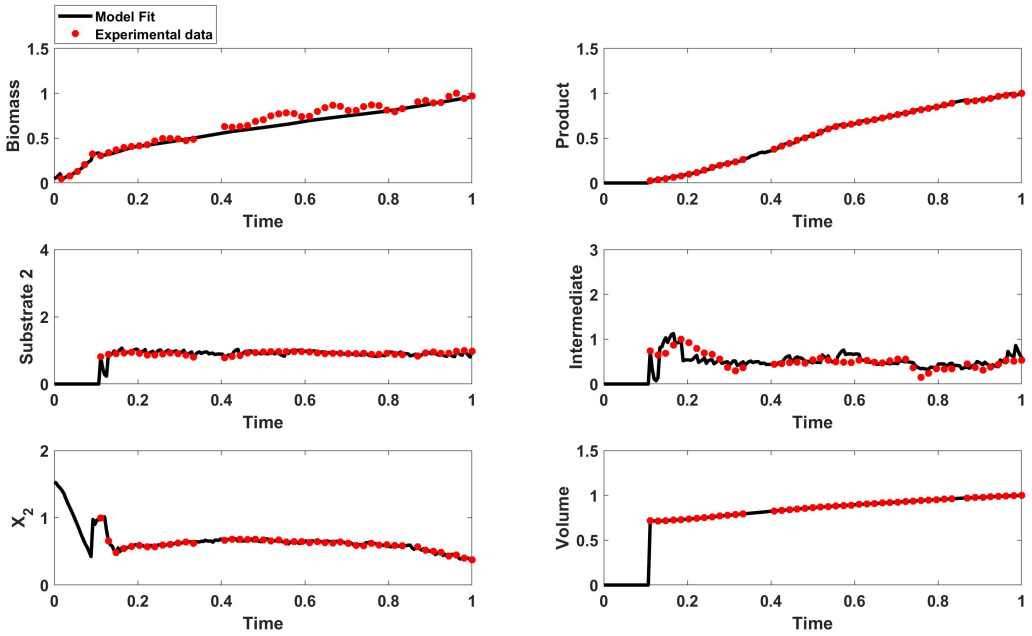

Figure 4.3: A comparison of the hybrid model-based observer fit and training data for phase 1 and 2 combined.

these targets and computing the optimal operating conditions of the bio-fermenter to maintain an optimal Substrate 2 concentration throughout the process. The practical considerations include setting bounds on temperature, flow rates, and rate of change of both of them. In order to do this, a Nonlinear program (NLP) is developed on GAMS.

According to the industry sponsor, it is desired to have a Substrate 2 concentration of $S_{2_{opt}}$, target product amount as $Prod_{opt}$ and primary unit as Eco_{opt}. Here, primary unit is described as follows:

$$Primary\ unit = \frac{Total\ Substrate\ 2\ used\ (l)}{Amount\ of\ Product\ (kg)} \tag{4.1}$$

The main target is to maximize product amount, i.e., get Product higher than $Prod_{opt}$, and minimize primary unit, i.e., get primary unit lesser than Eco_{opt}. To do that, an optimization problem is developed and it is given as follows:

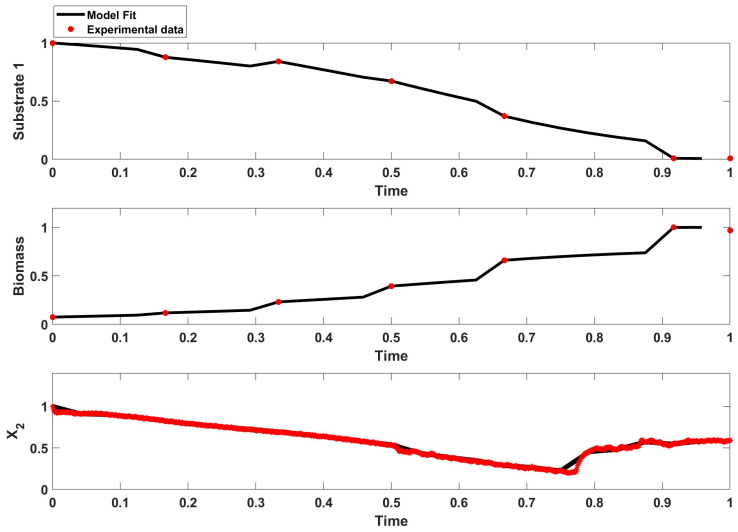

Figure 4.4: A comparison of the hybrid model-based observer fit and validation data during phase 1.

$$\min_{T,F_{in},X_1} \quad (Prod(f) - Prod_{opt})^2 - \frac{Prod_{opt}^2}{Eco_{opt}^2}(Eco(f) - Eco_{opt})^2 - 10^5 \times (TS - 195 \times S_{2_{opt}}) \tag{4.2a}$$

$$\text{s.t.} \quad 0.95 \times S_{2_{opt}} < S_2(t) < 1.05 \times S_{2_{opt}} \tag{4.2b}$$

$$78.8 < T < 88 \quad Temperature\ in\ F \tag{4.2c}$$

$$1500 < F_{in} < 1800 \quad Substrate\ 2\ flow\ rate\ in\ lb/h \tag{4.2d}$$

$$-50 < f_r < 50 \quad Rate\ of\ change\ of\ Substrate\ 2\ flow\ rate\ in\ l/h \tag{4.2e}$$

$$17 < X_1 < 25 \quad X_1\ Flow\ rate\ in\ lb/h \tag{4.2f}$$

$$-0.2 < t_r < 0.1 \quad Rate\ of\ change\ of\ Temperature\ in\ F \tag{4.2g}$$

where TS is the total sum of Substrate 2 concentrations over the duration of the process. The decision variables are Temperature T, Substrate 2 flow rate F_{in}, and X_1. A new decision variable f_r and t_r are introduced to denote the rate of change of Substrate 2 flow in l/h, and temperature, respectively. The product amount is given a free role to take any value, and as it is a maximization problem, the product tries to go as high as possible. There is a multiplication factor of $\frac{Prod_{opt}^2}{Eco_{opt}^2}$ in front of the primary unit term to equally weigh the primary unit. A term $10^5 \times (TS - 195 \times S_{2_{opt}})$

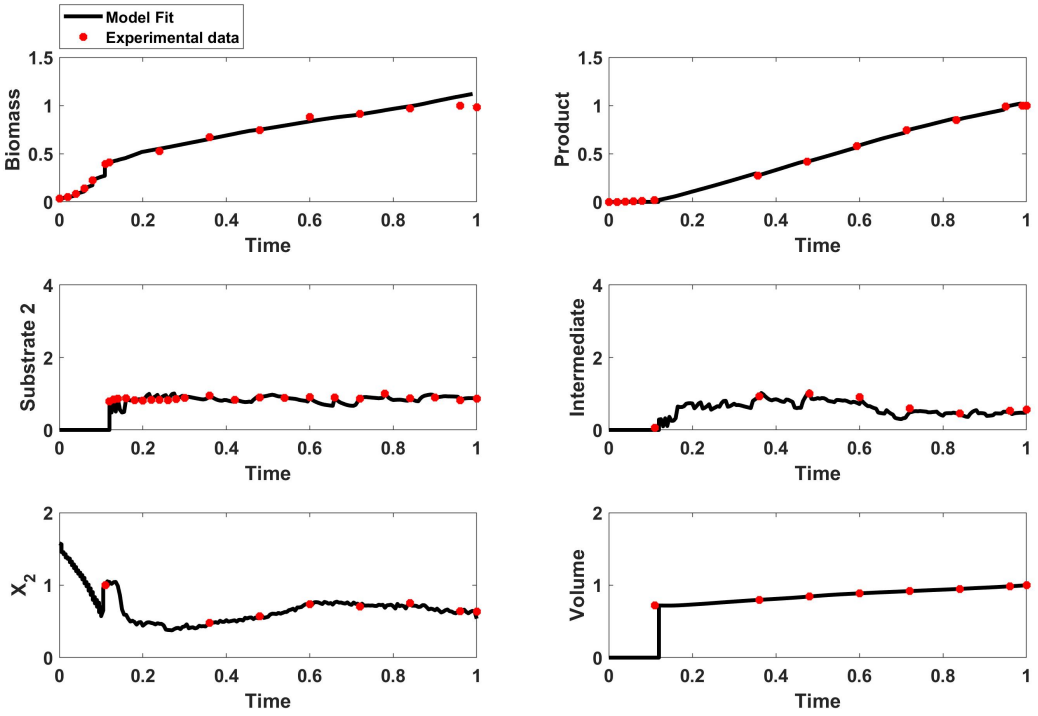

Figure 4.5: A comparison of the hybrid model-based observer fit and validation data for phase 1 and 2 combined.

is subtracted in order to maintain Substrate 2 concentration near $S_{2_{opt}}$. A factor of 10^5 is used so that the term in the bracket goes to near zero and the Substrate 2 concentration is driven to the desired value. Another important consideration is the volume end condition (Volume should be less than reactor volume of around 100,000 gallons) which is also the end point of the optimization problem.

The optimization problem is now solved for training dataset, and the optimal operating conditions and concentration of states are plotted and shown in Fig. 4.6, and Fig. 4.7, respectively. From the plots, it is seen that the modified control algorithm is able to maintain Substrate 2 concentration at $S_{2_{opt}}$ for most of the duration of the process and between $0.95 \times S_{2_{opt}}$- $1.05 \times S_{2_{opt}}$ throughout the process. Also, the product target is met and the product amount is found to be almost 10% more than the desired amount, and the volume end condition is satisfied too. The primary unit is found to be around $0.995 \times Eco_{opt}$ which is lower than the optimal target. The total x_1 is within the

bounds, and thus, the process is feasible. The input trajectories for temperature, Substrate 2 flow rate, and X_1 flow rate are seen to change dynamically with time throughout the process, and that plays a big role in maintaining the Substrate 2 concentration at $S_{2_{opt}}$. We also see that Intermediate, Biomass, and X_2 are also around the same range as the historical plant values thus confirming that the states are indeed optimal.

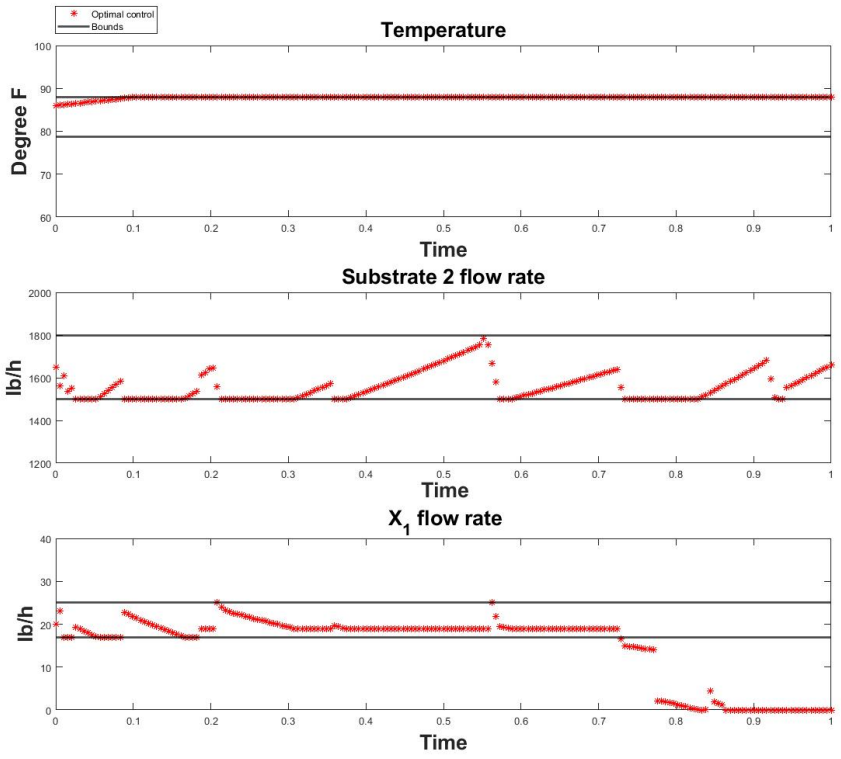

Figure 4.6: Optimal operating conditions and bounds for training data.

Now, the optimal control algorithm is implemented on the validation dataset, and the optimal operating conditions and concentration of states are plotted and shown in Fig. 4.8, and Fig. 4.9, respectively. It is important to note that this dataset runs for a lesser duration of time compared to the previous batch hence the product amount is lower than optimal target but it is still around 8% higher than the plant value. Also, the volume end condition is met and Substrate 2 is maintained within a 5% range of $S_{2_{opt}}$.

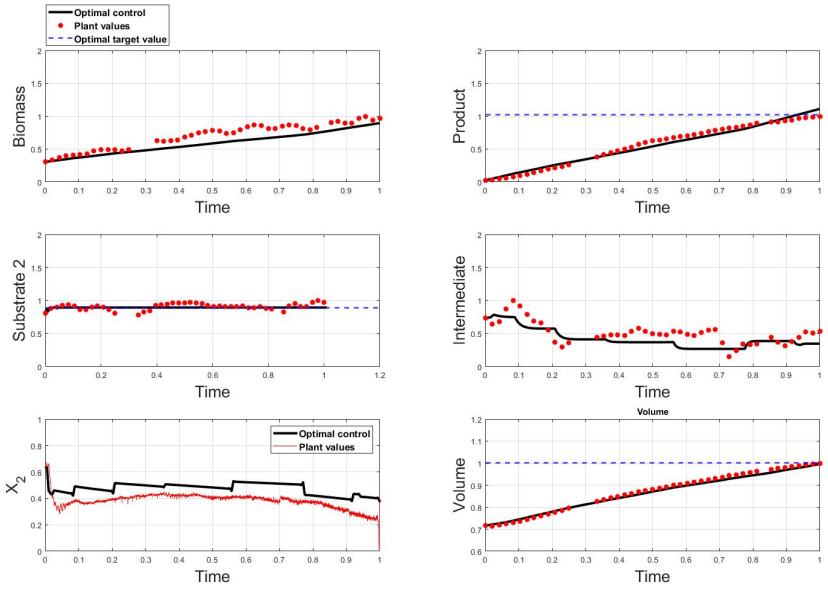

Figure 4.7: A comparison of optimal states, historical plant values, and optimal targets for training data.

Based on the above plots, we can say that the optimization problem is able to effectively maintain Substrate 2 concentration near the desired range throughout the process, while also meeting and improving upon the target product amount and primary unit. This control algorithm works well for a full-scale bio-fermentation process while considering practical temperature and flow rate constraints throughout the process. This control algorithm can now be used as part of a model predictive framework in order to implement the observer and optimization problem online for closed-loop operation.

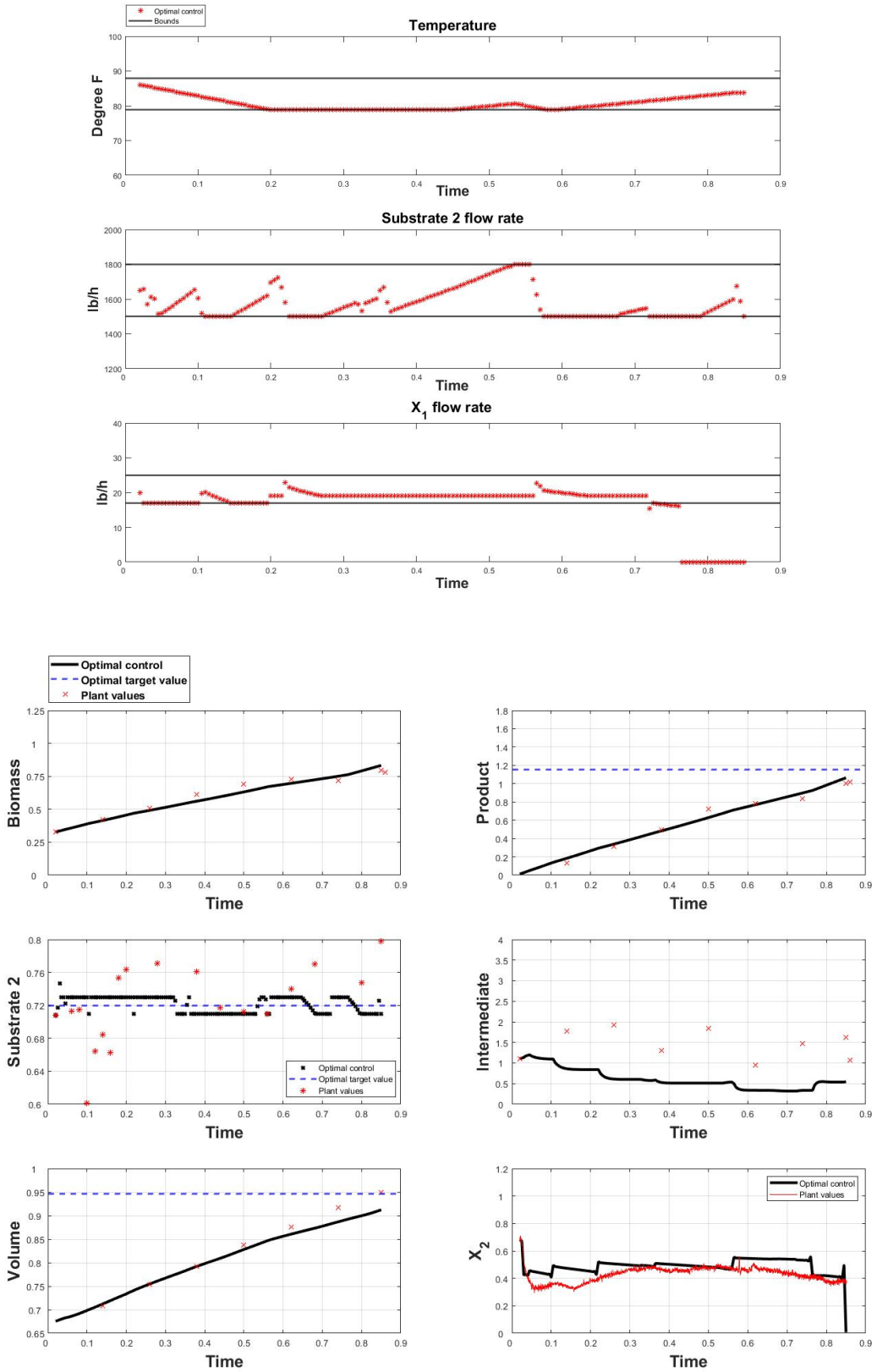

Figure 4.9: A comparison of optimal states, historical plant values, and optimal targets for validation data.

5. SUMMARY AND CONCLUSIONS

In this work a hybrid modeling strategy was proposed, where a DNN was used to predict uncertain process parameters in a bio-fermentation process. Due to a lack of first-principles approach to model the time-varying parameters, a sensitivity analysis was initially carried out to determine which model parameters had a huge influence on the model outputs. Then, a clustering approach was utilized to study the time sensitive nature of these model parameters, before employing those with large variations in the hybrid model. The hybrid model showed superior accuracy over the first-principles model, particularly for the estimation of Biomass, Product, and Substrate 2, which was of importance to the industry sponsor. This is attributed to the utilization of the DNN to accurately predict uncertain time-varying model parameters. Next, an open-loop hybrid model-based observer was designed using re-initialization of available measurements in order to accurately estimate the internal states of the process. All the states, especially Biomass, Substrate 2, and Product were tracked precisely compared to the available measurements within the sampled time, throughout the duration of the bio-fermentation process. This further reaffirms the prediction accuracy of the hybrid model. After designing the observer, an optimal control algorithm was developed which maximized the product amount and minimized the use of Substrate 2(responsible for cost of the process). Along with increasing productivity, a desired concentration of Substrate 2 was maintained while considering all the practical constraints like temperature, flow rate, and volume of the bio-fermenter. Thus, we were able to compute the optimal operating conditions of the bio-fermenter to maintain an optimal Substrate 2 concentration throughout the process while achieving optimal product and cost targets.

5.1 Further Study

Further research involves developing a model predictive control framework for closed-loop operation using the optimization problem used in order to perform online control of the bio-fermentation process. In order to do so, an interactive framework between MATLAB and GAMS

needs to be developed which imports and exports the manipulated inputs, real-time measurements, and optimal operating conditions of the process. In order to further improve the prediction accuracy of the states, a future direction can be to use adaptive hybrid modeling which is able to train the parameters while the process is running. This would ensure that accurate parameter estimation as well as further improve the observer and control algorithm accuracy.

REFERENCES

[1] L. Mears, S. M. Stocks, M. O. Albaek, G. Sin, and K. V. Gernaey, "Mechanistic fermentation models for process design, monitoring, and control.," *Trends in Biotechnology*, vol. 35, no. 10, pp. 914–924, 2017.

[2] B. Bhadriraju, M. S. F. Bangi, A. Narasingam, and J. S.-I. Kwon, "Operable adaptive sparse identification of systems: Application to chemical processes," *AIChE Journal*, vol. 66, no. 11, p. e16980, 2020.

[3] M. L. Thompson and M. A. Kramer, "Modeling chemical processes using prior knowledge and neural networks," *AIChE J.*, vol. 40, no. 8, pp. 1328–1340, 1994.

[4] D. C. Psichogios and L. H. Ungar, "A hybrid neural network-first principles approach to process modeling," *AIChE. J.*, vol. 38, no. 10, pp. 1499–1511, 1992.

[5] T. Bohlin and S. F. Graebe, "Issues in nonlinear stochastic grey box identification," *International Journal of Adaptive Control and Signal Processing*, vol. 9, no. 6, pp. 465–490, 1995.

[6] S. Gnoth, M. Jenzsch, R. Simutis, and A. Lübbert, "Product formation kinetics in genetically modified E. coli bacteria: inclusion body formation," *Bioprocess and Biosystems Engineering*, vol. 31, pp. 41–46, Jan 2008.

[7] G. Zahedi, A. Lohi, and K. Mahdi, "Hybrid modeling of ethylene to ethylene oxide heterogeneous reactor," *Fuel Processing Technology*, vol. 92, no. 9, pp. 1725 – 1732, 2011.

[8] S. Gupta, P.-H. Liu, S. A. Svoronos, R. Sharma, N. A. Abdek-Khalek, Y. Cheng, and H. El-Shall, "Hybrid first-principles/neural networks model for column flotation," *AIChE J.*, vol. 45, no. 3, pp. 557–566, 1999.

[9] P. Georgieva and S. de Azevedo, *Computational intelligence techniques for bioprocess modelling, supervision and control*. Springer, 2009.

[10] V. Mahalec and Y. Sanchez, "Inferential monitoring and optimization of crude separation units via hybrid models," *Comput. Chem. Eng.*, vol. 45, pp. 15 – 26, 2012.

[11] X. Wang, J. Chen, C. Liu, and F. Pan, "Hybrid modeling of penicillin fermentation process

based on least square support vector machine," *Chemical Engineering Research and Design*, vol. 88, no. 4, pp. 415 – 420, 2010.

[12] P.-C. Fu and J. Barford, "A hybrid neural network-first principles approach for modelling of cell metabolism," *Comput. Chem. Eng.*, vol. 20, no. 6, pp. 951 – 958, 1996.

[13] N. Carinhas, V. Bernal, A. P. Teixeira, M. J. T. Carrondo, P. M. Alves, and R. Oliveira, "Hybrid metabolic flux analysis: combining stoichiometric and statistical constraints to model the formation of complex recombinant products," *BMC Systems Biology*, vol. 5, no. 34, 2011.

[14] A. P. Teixeira, C. Alves, P. M. Alves, M. J. T. Corrondo, and R. Oliveira, "Hybrid elementary flux analysis/nonparametric modeling: application for bioprocess control," *BMC Bioinformatics*, vol. 8, no. 30, 2007.

[15] C. A. O. Nascimento, R. Giudici, and N. Scherbakoff, "Modeling of industrial nylon-6,6 polymerization process in a twin-screw extruder reactor. ii. neural networks and hybrid models," *Journal of Applied Polymer Science*, vol. 72, no. 7, pp. 905–912, 1999.

[16] M. Reuter, J. V. Deventer, and T. V. D. Walt, "A generalized neural-net kinetic rate equation," *Chem. Eng. Sci.*, vol. 48, no. 7, pp. 1281 – 1297, 1993.

[17] R. Jia, Z. Mao, Y. Chang, and L. Zhao, "Soft-sensor for copper extraction process in cobalt hydrometallurgy based on adaptive hybrid model," *Chemical Engineering Research and Design*, vol. 89, no. 6, pp. 722 – 728, 2011.

[18] H. C. Aguiar and R. M. Filho, "Neural network and hybrid model: a discussion about different modeling techniques to predict pulping degree with industrial data.," *Chemical Engineering Science*, vol. 56, no. 2, pp. 565 – 570, 2001.

[19] B. Fiedler and A. Schuppert, "Local identification of scalar hybrid models with tree structure," *IMA Journal of Applied Mathematics*, vol. 73, no. 3, pp. 449 – 476, 2008.

[20] M. R. Arahal, C. M. Cirre, and M. Berenguel, "Serial grey-box model of a stratified thermal tank for hierarchical control of a solar plant," *Solar Energy*, vol. 82, no. 5, pp. 441 – 451, 2008.

[21] J. Schubert, R. Simutis, M. Dors, I. Havlik, and A. Lübbert, "Bioprocess optimization and

control: Application of hybrid modelling," *Journal of Biotechnology*, vol. 35, no. 1, pp. 51 – 68, 1994.

[22] R. Eslamloueyan and P. Setoodeh, "Optimization of fed-batch recombinant yeast fermentation for ethanol production using a reduced dynamic flux balance model based on artificial neural networks," *Chemical Engineering Communications*, vol. 198, no. 11, pp. 1309–1338, 2011.

[23] M. S. F. Bangi and J. S.-I. Kwon, "Deep hybrid modeling of chemical process: Application to hydraulic fracturing," *Comput. Chem. Eng.*, vol. 134, p. 106696, 2020.

[24] D. Lee, A. Jayaraman, and J. S. Kwon, "Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling.," *PLOS Computational Biology*, vol. 16, no. 12, 2020.

[25] M. von Stosch, R. Oliveira, J. Peres, and S. F. de Azevedo, "Hybrid semi-parametric modeling in process systems engineering: Past, present and future," *Comput. Chem. Eng.*, vol. 60, pp. 86 – 101, 2014.

[26] D. Beluhan and S. Beluhan, "Hybrid modeling approach to on-line estimation of yeast biomass concentration in industrial bioreactor," *Biotechnology Letters*, vol. 22, pp. 631–635, 2000.

[27] R. G. Silva, A. J. Cruz, C. O. Hokka, R. L. Giordano, and R. C. Giordano, "A hybrid neural network algorithm for on-line state inference that accounts for differences in inoculum of cephalosporium acremonium in fed-batch fermentors," *Applied Biochemistry and Biotechnology.*, vol. 91-93, pp. 341–352, 2001.

[28] M. Ignova, G. C. Paul, C. A. Kent, C. R. Thomas, G. A. Montague, J. Glassey, and A. C. Ward, "Hybrid modelling for on-line penicillin fermentation optimisation," *IFAC Proceedings Volumes*, vol. 34, no. 1, pp. 395–400, 2002.

[29] A. J. M. Boareto, M. B. De Souza Jr, F. Valero, and B. Valdman, "A hybrid neural model (HNM) for the on-line monitoring of lipase production by candida rugosa," *Journal of Chemical Technology and Biotechnology*, vol. 82, no. 3, pp. 319–327, 2007.

[30] S. O. Laursen, D. Webb, and W. F. Ramirez, "Dynamic hybrid neural network model of an industrial fed-batch fermentation process to produce foreign protein," *Comput. Chem. Eng.*, vol. 31, no. 3, pp. 163–170, 2007.

[31] R. Eldan and O. Shamir, "The power of depth for feedforward neural networks," *Proceedings of the Twenty-Ninth Annual Conference on Learning Theory*, vol. 49, pp. 907–940, 2016.

[32] M. S. F. Bangi and J. S.-I. Kwon, "Deep reinforcement learning control of hydraulic fracturing," *Computers & Chemical Engineering*, vol. 154, p. 107489, 2021.

[33] E. A. Wan and R. Van Der Merwe, "The unscented kalman filter for nonlinear estimation," in *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No. 00EX373)*, pp. 153–158, Ieee, 2000.

[34] S. J. Julier and J. K. Uhlmann, "New extension of the kalman filter to nonlinear systems," in *Signal processing, sensor fusion, and target recognition VI*, vol. 3068, pp. 182–193, International Society for Optics and Photonics, 1997.

[35] G. Welch, G. Bishop, *et al.*, "An introduction to the kalman filter," 1995.

[36] C. Ling and C. Kravaris, "Multirate sampled-data observer design based on a continuous-time design," *IEEE Transactions on Automatic Control*, vol. 64, no. 12, pp. 5265–5272, 2019.

[37] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168, 1944.

[38] D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *SIAM J. Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[39] F.-S. Wang, T.-L. Su, and H.-J. Jang, "Hybrid differential evolution for problems of kinetic parameter estimation and dynamic optimization of an ethanol fermentation process," *Industrial & engineering chemistry research*, vol. 40, no. 13, pp. 2876–2885, 2001.

[40] S. Ö. Laursen, D. Webb, and W. F. Ramirez, "Dynamic hybrid neural network model of an industrial fed-batch fermentation process to produce foreign protein," *Computers & chemical engineering*, vol. 31, no. 3, pp. 163–170, 2007.

[41] M. C. Ordonez, J. P. Raftery, T. Jaladi, X. Chen, K. Kao, and M. N. Karim, "Modeling of

batch kinetics of aerobic carotenoid production using saccharomyces cerevisiae," *Biochemical Engineering Journal*, vol. 114, pp. 226–236, 2016.

[42] Z. Duan, T. Wilms, P. Neubauer, C. Kravaris, and M. N. C. Bournazou, "Model reduction of aerobic bioprocess models for efficient simulation," *Chemical Engineering Science*, vol. 217, p. 115512, 2020.

[43] Y. Chu and J. Hahn, "Parameter set selection for estimation of nonlinear dynamic systems," *AIChE J.*, vol. 53, no. 11, pp. 2858–2870, 2007.

[44] D. Lee, A. Singla, H.-J. Wu, and J. S.-I. Kwon, "An integrated numerical and experimental framework for modeling of CTB and GD1b ganglioside binding kinetics," *AIChE J.*, vol. 64, no. 11, pp. 3882–3893, 2018.

[45] B. Bhadriraju, A. Narasingam, and J. S.-I. Kwon, "Machine learning-based adaptive model identification of systems: Application to a chemical process," *Chemical Engineering Research and Design*, vol. 152, pp. 372–383, 2019.

[46] S. James, R. Legge, and H. Budman, "Comparative study of black-box and hybrid estimation methods in fed-batch fermentation," *Journal of process control*, vol. 12, no. 1, pp. 113–121, 2002.

[47] D. Lee, A. Jayaraman, and J. S. Kwon, "Development of a hybrid model for a partially known intracellular signaling pathway through correction term estimation and neural network modeling," *PLoS Computational Biology*, vol. 16, no. 12, p. e1008472, 2020.