Data-driven Models for Safety Management of Complex Systems

By

©Md. Alauddin

A thesis submitted to the School of Graduate Studies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Faculty of Engineering & Applied Science

Memorial University of Newfoundland

December 2021

St. John's   **Newfoundland**

Abstract

The advancement in big data and computing has prompted many industries including process industries to re-examine their traditional roles in design, control, and maintenance. The data-based models have the potential to contribute significantly to making the process industry more efficient, much safer, and environmentally friendlier. However, poor quality data, process uncertainty, random and spurious errors are major challenges to be handled by the data-driven methodologies. The objective of this research is to develop data-driven models for the safety management of complex process systems. This work also extends process safety principles to pandemic risk management to safeguard human health. Even though the Coronavirus Disease of 2019 pandemic has not resulted from process operations, energy industries were one the hardest hit sectors due to the pandemic. The COVID-19 pandemic has disrupted processing operations resulting in the historic collapse in demand and price crash leading to unprecedented scenarios. Nonetheless, the present pandemic provides numerous possibilities to strengthen engineering risk management approaches for benefiting all enterprises, especially, the oil and gas processing sectors. The reasons for studying pandemic risk management using the process safety framework are three-folds; to evince the multi-disciplinary nature of process safety principles, to demonstrate similarities between system safety and epidemiological risk management, and to manage pandemic risk using process safety principles.

This work presents robust data-based efficient methodologies for fault detection, fault characterization, and mitigation for ensuring safety. The robustness has been inculcated by explicitly addressing the data quality issues, reconciling data-driven models with mechanistic models, integrating meta-learning, and incorporating prior knowledge and expert opinions. This thesis addresses the data quality issue by developing a robust model based on harnessing data

quality features and assigning a lower weight to the low-quality data. The proposed method demonstrated improved results in detecting abnormalities in two case studies; a continuous stirred tank heater problem and the Tennessee Eastman (TE) benchmark chemical process. The percentage improvements in accuracy in detecting the step fault (IDV-1 fault) of the TE process were 1.0 %, and 4.5 % on 1%, and 10% mislabeled data respectively. A process dynamics-guided neural network model has been proposed to improve generalization. This has been implemented by adding an additional layer to the deep neural network architecture to incorporate process dynamics such as material and energy balance equations, universal laws, standard correlations, and field knowledge. The proposal has been evaluated on regression and classification tasks representing transient and steady-state operations of chemical processing systems. It resulted in significant gains in predicting dependencies in steady-state operations and forecasting transient conditions due to its improved generalization ability. The fault detection of a neural network model has been improved by incorporating meta-learning as well. The thesis presents a data-driven fault detection model using an artificial neural network and variable mosquito flying optimization technique for parameter tuning by maximizing fault detection rate while minimizing the false alarm rate. The proposed fault detection method has been applied for detecting faults in the TE benchmark process.

The thesis also presents a hybrid formalism in pandemic risk management where the performance of the mechanistic models has been improved with advanced data-driven approaches. Thus, an artificial neural network-based susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) model has been devised to effectively capture the temporal variability of a disease spread. It also addresses inconsistencies in reporting of infected cases by assigning a higher weight to the mortality data which is a more credible indicator of the disease progression than the reported

infected cases. This yielded satisfactory results in forecasting infection cases of the COVID-19 in Ontario, British Columbia, Italy, and Germany. Uncertainty is another critical factor of processing systems and epidemiological modeling. This thesis quantifies the risk with randomness in the model parameters such as incubation, infection, and recovery periods under distinct measures of lockdown, schools and business closures including no measure. The uncertainty caused due to government interventions, changes in individual behavior, and the advent of multiple waves of an outbreak has been accounted by the parameter sharing feature of a hierarchical Bayesian network. Markov chain Monte Carlo simulation has been used to study the variability in the hierarchical Bayesian network. Finally, the thesis presents pandemic risk management frameworks using engineering safety principles such as precautionary, as low as reasonably practicable, and the layer of protection analysis approaches. An event tree model of pandemic risk management for distinct risk-reducing strategies realized due to natural evolution, government interventions, societal responses, and individual practices has been proposed. The proposed framework also investigates the impacts of distinct interventions on the survivability of an infected individual under existing healthcare facilities.

This thesis can help in ensuring safety of complex systems by detecting abnormalities using robust data-driven and semi-mechanistic models. The thesis explores the synergy between process safety and epidemiology to better understand, analyze, and manage the risk. This is a first step towards the interdisciplinary study of this sort and paves the way for more interdisciplinary studies, enriching both disciplines.

Acknowledgment

This work would have not been completed without the extensive support from many individuals and organizations. I am very grateful to them.

Table of Contents

List of Figures

List of Tables

Authorship Statement

This thesis is a manuscript format thesis. Each chapter of the thesis is either published or in preparation for publication in a peer-reviewed journal. I, Md. Alauddin, hold a principle author status for all the manuscript chapters (Chapter 3 - 8) in this thesis. However, each manuscript is co-authored by my supervisor, co-supervisors, and collaborators who contributed to conceptualization, development of this work, writing, review & editing, supervision, project administration, and funding acquisition. The bibliographic citation and contributions of the authors in each chapter are presented below:

Chapter-3

Alauddin, M., Khan, F., Imtiaz, S., Ahmad, S., & Amyotte, P., Quality-based training (QbT) of supervised neural networks for robust fault detection in the presence of mislabeled data, *(draft under the committee review)*

> **Md Alauddin:** Formal Analysis, Methodology, Software; Investigation, Validation; Writing - Original Draft; Writing - Review & Editing
>
> **Faisal Khan:** Conceptualization, Methodology, Writing - Review & Editing; Supervision; Project administration; Funding acquisition
>
> **Syed Imtiaz:** Conceptualization, Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition
>
> **Salim Ahmed:** Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition
>
> **Paul Amyotte:** Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition

Chapter-4

Alauddin, M., Khan, F., Imtiaz, S., Ahmad, S., & Amyotte, P., A primer on integrating process dynamics in data-driven models of chemical processing systems, *(draft under committee review)*

As a first author, I developed the modeling of the system, analyzed the results and prepared the first draft of the manuscript. Co-authors, Drs. Khan, Imtiaz, Ahmed, and Paul conceptualize the problem, provided financial and technical support in conducting the experiments and helped in analyzing the data, and also helped in revising the manuscript. They also provided technical guidance and also provided valuable feedback in improving the manuscript.

Chapter-5

Alauddin, M., Khan, F., Imtiaz, S., & Ahmad, S. (2019). A variable mosquito flying optimization-based hybrid ANN model for process fault detection systems. *Process Safety Progress, 39 (1), 1-8.*

I (Md Alauddin) was the primary author and responsible for the problem formulation, modeling, analysis, and preparation of the manuscript. Drs. Khan, Imtiaz, and Ahmed contributed to the methodologies, development, and refinement of the presentation.

Chapter-6

Alauddin, M., Khan, A., Khan, F., Imtiaz, S., Ahmad, S., & Amyotte, P., (2020). How Can Process Safety and a Risk Management Approach Guide Pandemic Risk Management*? Journal of Loss Prevention in Process Industries,* 68, 104310.

As a primary author, I (Md Alauddin) was responsible for modeling, analysis, and writing of this manuscript. Dr. Amin contributed to the writing and presentation of this work. Dr. Khan was instrumental in conceptualization, developing methodology, writing, review & editing, administration, and funding acquisition. Drs. Imtiaz, Ahmed, and Paul contributed to developing methodology, validation, formal analysis, writing, review & editing, supervision, and funding acquisition.

Chapter-7

Alauddin, M., Khan, F., Imtiaz, S., Ahmad, S., Amyotte, P., & Vanberkel, P. (2021). Pandemic

Risk Assessment and Management in Bayesian Framework. *(submitted to Risk Analysis Journal).*

> This was a collaborative work among the Centre for Risk, Integrity, and Safety Engineering
> (C-RISE), Faculty of Engineering and Applied Science, Memorial University of
> Newfoundland, Department of Process Engineering and Applied Science, Dalhousie
> University, and Department of Industrial Engineering, Dalhousie University. I (Md
> Alauddin) was the primary author and all the co-authors helped in developing the
> manuscript. The contributions of the author and co-author are as follows-
>
> **Md Alauddin:** Formal Analysis, Methodology, Software; Investigation, Validation;
> Writing - Original Draft; Writing - Review & Editing
>
> **Faisal Khan:** Conceptualization, Methodology, Writing - Review & Editing; Supervision;
> Project administration; Funding acquisition
>
> **Syed Imtiaz:** Methodology, Validation; Formal Analysis; Writing - Review & Editing;
> Supervision; Funding acquisition
>
> **Salim Ahmed:** Methodology, Validation; Formal Analysis; Writing - Review & Editing;
> Supervision; Funding acquisition
>
> **Paul Amyotte:** Methodology, Validation; Formal Analysis; Writing - Review & Editing;
> Supervision; Funding acquisition
>
> **Peter Vanberkel:** Methodology, Validation; Formal Analysis; Writing - Review &
> Editing; Supervision; Funding acquisition

Chapter-8

Alauddin, M., Khan, F., Imtiaz, S., Ahmad, S., & Amyotte, P., (2021). Pandemic Risk

Management using Engineering Safety Principles. *Process Safety and Environmental Protection*,

150, 416-432.

> I was the primary author and contributed to conceptualization, modeling, analysis,
> manuscript preparation, and presentation. The co-authors (2-5) contributed to
> conceptualization, formal analysis, manuscript development, refinement, and presentation.

Chapter 1

Introduction

1.1 Background and Motivation

Chemical process systems are high volume, hazardous operations that can result in significant loss of lives and assets if an accident occurs. Ensuring efficient process monitoring and early detection and diagnosis of abnormalities is essential to prevent such losses. This can be achieved by rendering preventive actions such as process monitoring, system reconfiguration, installing safety instrumented systems, and facilitating timely maintenance (Niu, Yang, & Pecht, 2010). The enhanced design and technological advancements of the modern era have led to a decline in the number of catastrophic accidents. Nonetheless, accidents in process industries continue to occur (Khan & Abbasi, 1999; Mannan, Reyes-Valdes, Jain, Tamim, & Ahammad, 2016). Thus, the current safety technologies need to be continuously reviewed and updated to prevent adverse incidents.

Efficient detection and diagnosis of the abnormalities at an early stage are of utmost importance in ensuring operational safety in process systems. Early detection of abnormal conditions gives a lead time to respond effectively to prevent or mitigate the damage from the abnormal condition. Thus, robust techniques of fault detection, identification, and diagnosis have become vital tools for the safety management of process systems. Many fault detection and diagnostic (FDD) methods have been developed over the last four decades. They can broadly be divided in two categories: model-based and data-based (Alauddin, Khan, Imtiaz, & Ahmed, 2018; Venkatasubramanian, Rengaswamy, Yin, & Kavuri, 2003). Experts predict that the 21st century will be known for the development of data-based systems that are revolutionizing the process design and monitoring of industrial systems. Many oil and gas companies are looking forward to

adopting artificial intelligence-based models for their operations. According to a whitepaper by the World Economic Forum, "digitalization in the oil and gas sector has the potential to unlock around $1.6 TN of value over the next decade (World Economic Forum & Accenture, 2017)".

There is a growing consensus towards intelligent and data-driven operations for design and predictive maintenance in processing systems. However, some of the underlying factors must be considered while devising methodologies for the safety management of complex process systems.

i.   *The complex nature of the system:* Chemical process systems are becoming extremely complex due to increased automation and process intensification in the pursuit of improved efficiency. A large number of associated variables and system parameters cause difficulty in efficient monitoring and safety management of complex industrial systems. For instance, an average of 14,250 alarms, with a peak of 26,650 alarms per day, was recorded in a European refinery (Ren, Zhu, Cai, & Li, 2017). Redundant alarms are usually being ignored by operators. Consequently, alarm flooding is becoming a threat to complex industrial systems. The highly correlated nature of the multivariate data is another challenge to be dealt with for effective monitoring and ensuring safety of process systems. Thus, efficient techniques of process monitoring, fault detection, identification, and diagnosis are needed.

ii.   *The dynamic nature of process operations:* The dynamic nature of process operations is another challenging task to be taken into account for the effective management of incipient abnormalities. Disturbances, transportation lag, and process dynamics constitute the dynamic behavior of process systems. Process controls are used to respond to the myriad disturbances in processing facilities. A process safety management system should be capable of identifying these abnormal situations and taking mitigation actions for handling these disturbances.

iii.   *Uncertainty in process systems:* Uncertainty in data is a common characteristic of most process industries. The common sources of uncertainty include measurement methods, manufacturers' imprecise specification, calibration, data acquisition and processing, and malfunctioning of secondary devices. Poor repeatability i.e., "the difference between two successive results obtained by the same operator with the same apparatus under constant operating conditions" and reproducibility, "the difference between two results obtained by different operators in different laboratories on identical test materials" are other contributing factors to uncertainty in process systems. Inappropriate sampling techniques, improper sample conditioning, collecting samples from non-representative locations and non-representative operating conditions are also common sources of uncertainty in process systems. Process system uncertainty is further induced by randomness, spurious errors, and malfunctioning of equipment.

Uncertainty can be minimized by following appropriate sampling, measurement, estimate, and quality assurance methods, but it cannot be eliminated. Thus, robust data-driven techniques for handling uncertainties in the safety management of processing systems are required.

iv.   *Inconsistent and sparse data:* Process systems are going through digital transformations and data-based management systems. The success of data-based models is utterly susceptible to the nature and availability of training data. Sparse data are also a growing concern for the efficient training of data-driven models. Usually, a large volume of data is available to represent the normal operating conditions, while faulty condition data is relatively scarce in industrial systems. The imbalanced data can lead to a biased outcome if not addressed in the modeling. Accordingly, robust models are required for dealing with inconsistent and sparse data of industrial systems.

v.   *Natural hazards and unseen events:* Earthquakes, floods, cyclones, pandemic outbreaks, and other natural catastrophes continue to pose a threat to the safety of process systems. Natural-

hazard-triggered technological events can exacerbate the impact of a natural disaster. It can be detrimental to health, livelihood, social system, and the environment. The COVID-19 pandemic, for example, disrupted processing operations resulting in the historic collapse in demand and price leading to unprecedented scenarios (Camp, Mead, Reed, Sitter, & Wasilewski, 2020; OECD, 2020). Personal safety and operability of process systems were jeopardized as a result. Efficient methods for combating these outbreaks need to be devised.

The epidemiological modeling confronts similar challenges as outlined in the preceding section. The trajectory of an epidemic disease depends on the imposed regulations, societal response, individual practices, and the advent of multiple waves of disease. Uncertainty and inconsistency are inevitable in epidemiological studies. A pandemic modeling comprises both aleatory uncertainty (caused by variability in population/data) and epistemic uncertainty (arising from a lack of knowledge of the phenomena). The distinct strains of the virus, modes of propagation (airborne or contact transmission), and uncertainty in infectivity, rate of incubation, infection, and recovery all constitute uncertainty in epidemic modeling. Moreover, the reporting of the infected cases is often flawed by numerous factors such as lack of systematic testing, inherent delay between the date that an illness starts and the date the case is reported to public health authorities. These factors can significantly affect the outcomes of the risk management efforts if not considered in the models.

Data-driven, also known as empirical models, are generic and can adapt to meet the requirements of the modern era's complex systems. Data-based models have the potential to significantly improve efficiency, safety, and sustainability of process systems. The data-based models, especially the artificial neural network, can approximate complex functions. Machine learning models can be very good at fitting observations yet predictions may be contradictory with the laws

of physics or even physically nonsensical (e.g. negative density estimations) (Svendsen et al., 2021). Data-based models assume that the data is of sufficient granularity and quantity. They also suffer from the lack of generalization ability and struggle when data is scarce or in extrapolation regimes (Fisher et al., 2020; Pawar, San, Aksoylu, Rasheed, & Kvamsdal, 2021). Moreover, the "black box" nature of the data-driven models, especially deep neural networks, is a barrier to large-scale adoption in industrial systems. The data-based models can be improved by conforming mechanistic science, expert knowledge, and meta-learning insights.

First principle models, on the other hand, are robust and have a higher degree of generalization. Their system understanding feature compensates for the lack of high-quality data (Fisher et al., 2020). Developing efficient mechanistic models is often challenging and expensive, especially for new application areas with limited domain knowledge. Simplified first principle-based models are inapt to adequately capture the behavior of complex processing systems.

The objective of this work is to develop robust data-driven models for the safety of complex systems and human health management. The objective is achieved through three sub-objectives, to develop robust data-driven models for the safety of process systems, to devise semi-mechanistic models for assessing pandemic risk, and to establish synergy between process safety and pandemic risk management.

a) *The development of robust data-driven models for the safety of process systems:* Robust data-based models based on efficient algorithms, recombination, and hybridization of state-of-the-art methodologies have been devised for the abnormal situation management of processing facilities. The thesis presents robust data-driven models by exploiting data quality, reconciling data-driven models with mechanistic modeling, integrating meta-learning, and incorporating prior knowledge and expert opinions.

b) *Devising semi-mechanistic models for assessing pandemic risk:* The thesis presents integrating data-driven models in mechanistic frameworks for assessing pandemic risk under evolving conditions. This work shows how the pandemic risk forecast of a mechanistic model can be made more precise and conclusive using artificial neural networks and Bayesian formalisms.

c) *To establish synergy between process safety and pandemic risk management:* This thesis explores the synergy between process safety principles and epidemiology to better understand, analyze, and manage the risk. The pandemic risk has been analyzed using process safety principles such as precautionary principles, as low as reasonably practicable approach, event tree model, and the layer of protection analysis.

The thesis presents robust models for the safety management of complex processing systems and pandemic risk management (Fig. 1.1). This has been carried out by devising efficient methodologies for fault detection, fault characterization, and mitigation. The robustness has been instilled by exploiting data quality, reconciling data-driven models with mechanistic modeling, integrating meta-learning, and incorporating prior knowledge and expert opinions. Poor quality, sparsity, and inconsistency of data have also been effectively addressed by multiple formulations (Fig. 1.2). The hybrid formulations of data-based models and science-based models have been used in pandemic risk management as well. The thesis addresses the aforementioned challenges of process systems and pandemic risk management as follows:

i. *Handling low-quality data:* This thesis addresses the data quality issue by developing a robust model by exploiting data quality in the training of supervised networks (Chapter 3). Low-quality is also a common feature for pandemic data. The early growth dynamics of an epidemic is a better characterizing of the impact of an outbreak. Identifying significant features of the

initial growth kinetics of an outbreak is useful in reliable forecast and effective management of a pandemic. However, the newly infected-cases data are often flawed by numerous factors such as lack of systematic testing, the inherent delay between the onset of the disease, and the date that the case is reported to public health authorities. Mortality data are lagged but a more reliable indicator of the disease progression across populations than the number of confirmed cases. The data of the initial phase of the pandemic have been recalibrated based on first mortality data and a higher weight is assigned to more reliable data in fitting the model (Chapter 6).

ii. *Handling uncertainties in the system:* Uncertainty is another critical factor of processing systems and epidemiological modeling. This thesis addresses this issue by contributing as follows-

   a. *Monte Carlo simulation-based risk assessment framework:* Monte Carlo simulation has been employed to capture randomness in the model parameters (Chapter 6 and Chapter 8). The distributions of the model parameter such as incubation, infection, and recovery periods have been used to capture the long tail of the infection risk (Section 6.3.2). It has also been used to study the effect of enforcing and relaxing interventions for managing a pandemic outbreak (Section 6.3.3 and Section 6.3.4). Monte Carlo simulation has been used in assessing pandemic risk using Paté-Cornell's six levels of uncertainty (Section 8.3.1). Uncertainty has also been quantized in managing risk using precautionary and ALARP approaches (Section 8.3.1 and Section 8.3.3) and survival analysis of an infected individual under existing healthcare facilities (Section 8.3.4).

   b. *Parameter sharing feature of the hierarchical Bayesian structure*: The uncertainty in the reported pandemic data has been handled using the parameter sharing feature of the

hierarchical Bayesian structure. Markov chain Monte Carlo simulation (MCMC) has been used to study the variability in a hierarchical Bayesian framework (Section 7.3.2).

c. *Impact analysis for random parameters using a Bayesian framework:* A Bayesian inference network has been used to analyze the impact of distinct parameters in pandemic risk management. The thesis presents relevance-based reasoning, which is based on the conditional probability distribution, for impact analysis of availing the existing health facilities (Section 7.4.2).

iii. *Integrating system knowledge in the data-driven model:* Many methods have been presented in improving the performance of data-driven methods by optimizing hyper-parameters such as the number of hidden layers, number of neurons in each layer, the learning rate, and optimization techniques. Numerous approaches such as early stopping, regularization, dropout, and retraining have been devised for handling the overfitting and under-fitting of data-based models. However, there is a lack of literature on integrating the process knowledge such as material and energy balances, physical laws, chemical kinetics, and expert knowledge in the training of data-based models. Similarly, infectious disease dynamics have been extensively studied using data-driven and mechanistic models. The time series analysis of pandemic data has been carried out using numerous data-driven approaches i.e., autoregressive integrated moving average (ARIMA) technique, multivariate ARIMA, autoregressive moving average with exogenous inputs (ARMAX), and exponential smoothing (ETS). Advanced machine learning techniques such as artificial neural networks, deep neural networks, and recurrent networks such as long short-term memory (LSTM) have been used for time series analyses of infectious diseases (Zhang et al., 2013) (Volkova, Ayton, Porterfield, & Corley, 2017; Zhu et al., 2019). Several mechanistic models comprise compartmental, spatial, meta-population,

network-based, and agent-based methods have been developed. However, integrating the data-driven and mechanistic models is barely discussed.

This thesis presents a process dynamics-guided neural network (PDNN) model for improving performance of the system (Chapter 4). Similarly, a meta-learning-based hybrid model has been developed for refining the model parameters for an intended function (Chapter 5). Integrating data-based and science-based models have been enacted in pandemic risk management as well. The thesis presents a susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) model that can account for the asymptomatic transmission and hospitalization cases. The forecast has been made more compelling by capturing temporal changes in parameters of the SEIQRD model using an artificial neural network (Chapter 6) and a hierarchical Bayesian formalism (Chapter 7).

.

Fig. 1.1: Operational safety management using data-driven models of fault detection, characterization, and mitigation
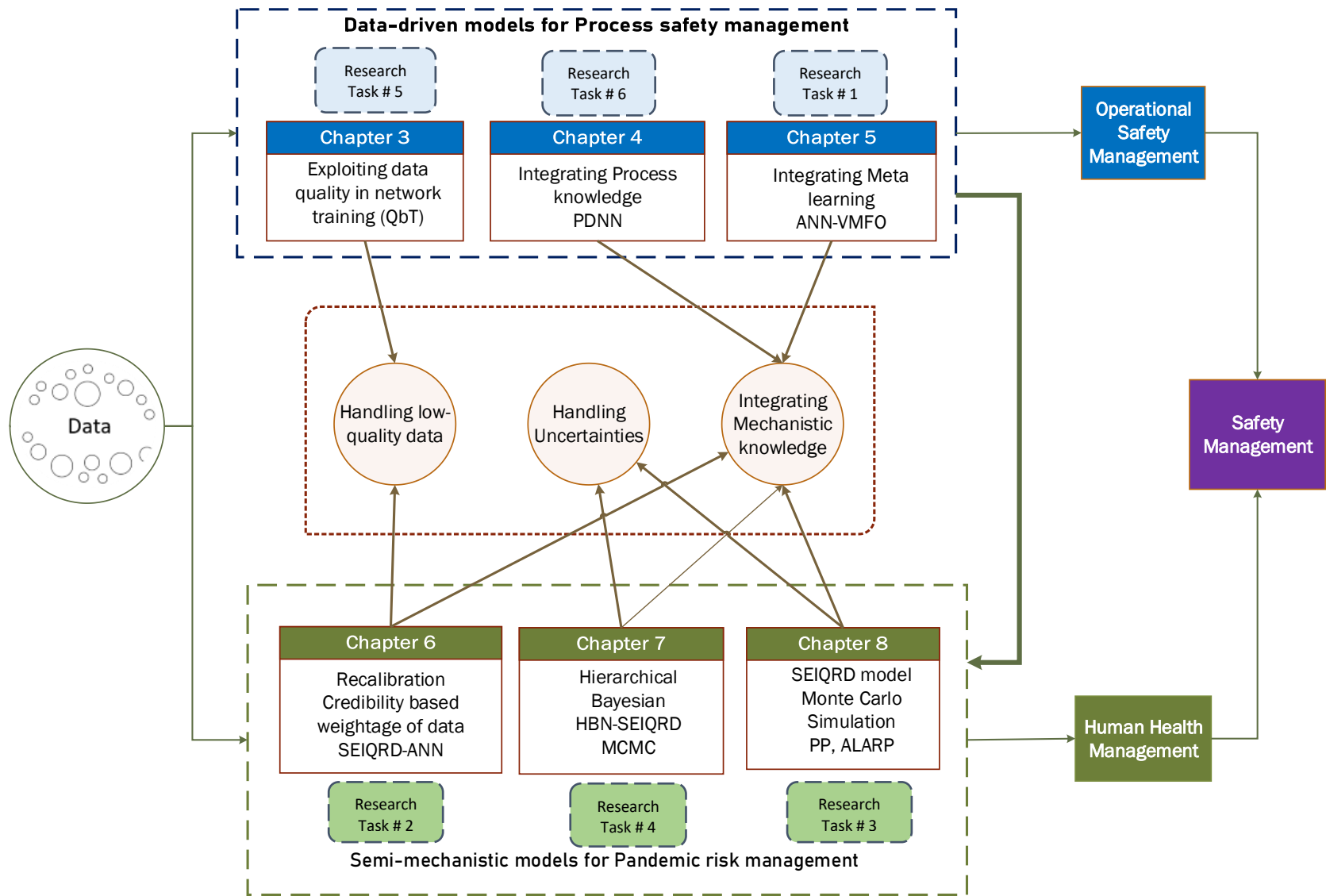
Fig. 1.2: Development of robust data-driven models for process safety and human health management

1.2 Contributions

This section summarizes the novelty and contributions of this research. The salient methodological and applicative contributions are listed below.

i. Devising a robust deep neural network model for handling low quality and mislabeled data

Application – The proposed model was examined for fault detection and isolation on two case studies; a continuous stirred tank heater problem and the Tennessee Eastman chemical process.

ii. Formulation and design of process-informed neural network model for abnormal situation management by integrating process knowledge with data.

Application – The proposed model has been examined on regression and classification tasks of the processing system representing steady-state and transient operations. It also resulted in improved generalization ability on the unseen and reduced sample-sized training data.

iii. Development of hybrid fault detection and isolation model using an artificial neural network (ANN) and variable mosquito flying optimization (V-MFO) technique.

Application – The proposed model has been examined on the Tennessee Eastman benchmark process for detecting abnormalities. The parameter tuning using the variable mosquito flying optimization demonstrated significant improvement in missed detection rate compared to the simple ANN.

iv. Development of a novel advanced semi-mechanistic SEIQRD pandemic model based on artificial neural network

Application – The proposed model has been used to assess the infection risk at distinct locations due to the COVID-19. This also investigates the risk in enforcing and releasing distinct risk-reducing measures including no measure.

v.   Proposed hierarchical variational Bayesian network-based semi-mechanistic model (HBN-SEIQRD) for capturing variability in parameter learning under evolving conditions.

Application – The proposed model has been used to assess the infection risk at distinct locations due to the multiple waves of the COVID-19 outbreak. This also investigates the impact of the COVID-19 outbreak on peak hospitalization cases under three scenarios: no measures or few interventions, moderate interventions, and stricter interventions.

vi.   Devising pandemic risk management frameworks using engineering safety principles such as precautionary, ALARP, and layer of protection analysis approaches

Application – Engineering safety principles-based pandemic risk managing formulations have been studied on the COVID-19 pandemic. The qualitative and quantitative analyses have been presented using the precautionary principle, ALARP approach, event tree analysis, and layer of protection analysis.

A significant contribution of this thesis is to explore the similarities between process safety and epidemiology to better understand, analyze, and manage the risk. Approaches to prevent, control, and mitigate infection are analogous to the hazard control and safety frameworks used in the process industries. We observed how early detection of abnormalities, quick response, effective monitoring, and enforcement to restrict the COVID-19 spread were effective in minimizing the pandemic risk. By drawing a parallel between two domains, it can be believed that the lessons learned from the COVID-19 pandemic would immensely benefit engineering safety personnel and healthcare experts in efficient policymaking.

1.3 Thesis Outline

The thesis is divided into nine chapters. The first few chapters are focusing on introducing the process safety concepts, data-based methodologies, data quality issues, and optimization methods

to calibrate process models. The semi-mechanistic models of pandemic risk management have been discussed in the following chapters. The last chapter presents the applications of process safety principles to pandemic risk management.

The first chapter outlines the problem, objectives, and scopes of this study. Chapter 2 presents the literature review related to advanced data-driven models for process safety management. The review focuses mainly on multivariate models including principal component analysis, partial least squares, independent component analysis, artificial neural networks, mixture models, and Bayesian approaches. This chapter also includes a review on modeling infectious disease transmissions with a special emphasis on dynamics and control of an outbreak. This section also discusses distinct approaches to estimate model parameters. The technical sections of the thesis have been depicted in Fig. 1.3.

Chapter 3 presents a novel approach, the quality-based training (QbT) for handling mislabeled and/or low-quality data in a deep neural network. In this Chapter, a new metric, the classification index, was formulated for effective evaluation of fault detection system. It demonstrates how the proposed robust ANN model overcomes the limitation of simple ANN in presence of low-quality data.

Chapter 4 presents a process dynamics-guided neural network model to improve the model generalization by rendering process dynamics and field expertise. It also presents an evaluation of the proposed models on regression and classification tasks of processing systems. A similar approach to integrate meta-learning in the deep neural network has been presented in Chapter 5. A hybrid model based on an artificial neural network and variable mosquito flying optimization technique has been used to detect faults based on maximizing fault detection rate and minimizing false alarm rate.

Chapter 6 presents an advanced semi-mechanistic model elicited from a compartmental model and an artificial neural network. The model has been developed to forecast the dynamic risk under evolving conditions. Monte Carlo simulation has been used to capture randomness of the model parameters. The risk in enforcement and releasing of distinct non-pharmaceutical interventions at different stages of a pandemic has also been evaluated in this Chapter. Chapter 6 also formulates a layer of protection analysis (LOPA) for preventive, controlling, and mitigating strategies for pandemic risk. Also, several areas of similarities have been identified where process safety and epidemiology can be benefited from each other.

Chapter 7 presents a Bayesian-based semi-mechanistic formalism for a credible assessment of pandemic risk. The variability caused by evolving conditions due to imposed regulations, varied individual responses, and the advent of the multiple waves of a pandemic outbreak has been captured using a hierarchical Bayesian network. This has been used to estimate parameters of the susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) model to forecast newly infected cases and fatality caused by the COVID-19. This formalism is a righteous demonstration of handling uncertainties using integrated knowledge where the semi-mechanistic model is enriched with mechanistic knowledge in the form of a set of differential equations of the disease dynamics, prior knowledge of parameter distribution, and pooling information from distinct periods of the pandemic outbreaks.

The relevancy of engineering safety principles to pandemic risk management has been introduced in Chapter 8. It assesses risk management strategies such as precautionary and as low as reasonably applicable (ALARP) approaches. The pandemic impact has been quantified using the susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) mechanistic model while the uncertainties have been accounted for using Monte Carlo simulation and Paté-Cornell's approach.

This Chapter also categorizes distinct risk-reducing strategies into hierarchical safety frameworks. Finally, an event tree model of pandemic risk management for distinct risk-reducing strategies realized due to natural evolution, government interventions, societal responses, and individual practices has been proposed. It also investigates the roles of distinct interventions on survivability of an infected individual under existing healthcare facilities. Chapter 9 summarizes the findings and practical implications of the study and identifies areas for further research. The overall structure of the thesis along with the tools for the intended tasks is presented in Table 1.1.
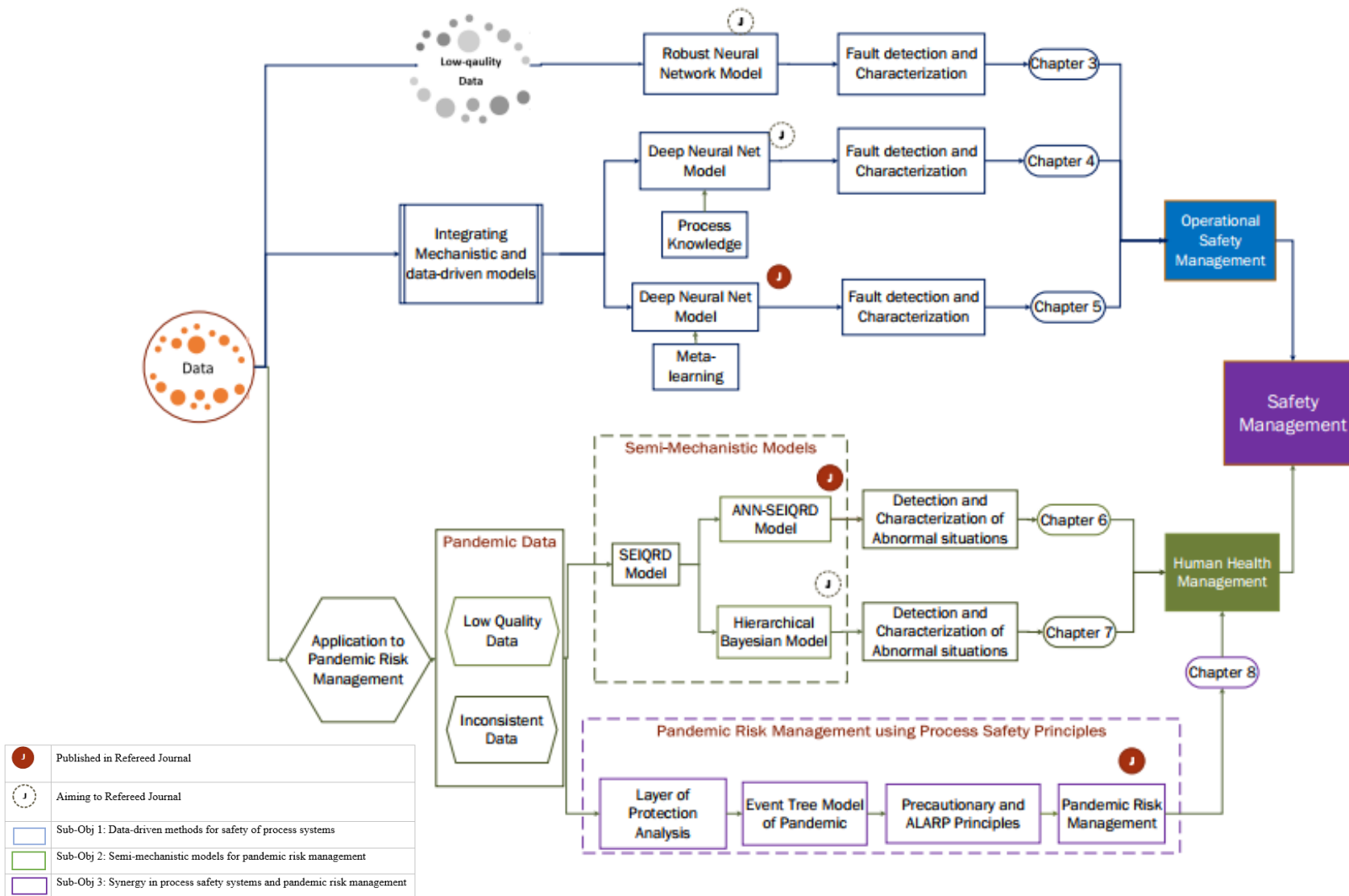
Fig. 1.3: The structure of the thesis

Table 1.1: Contribution of Chapters to the research objectives and associated tasks

Chapters and Titles	Research Objective	Associated tasks/ handling issues	Tools
Chapter 3: Quality-based training (QbT) of supervised neural networks for the robust fault detection in the presence of mislabeled data	*To develop robust data-driven models for the safety of process systems*	• addresses the data-quality issue by developing a robust model by exploiting data quality	Python 3.6 pandas pyplot
Chapter 4: A primer on integrating process dynamics in data-driven models of chemical processing systems	*To develop robust data-driven models for the safety of process systems*	• Integrating process dynamics in the data-driven model • Handle generalization of data-driven models	Python 3.6 Julia 1.62 SciML odenet
Chapter 5: A variable mosquito flying optimization based hybrid ANN model for fault detection of process systems	*To develop robust data-driven models for the safety of process systems*	• Integrating meta-learning-based in the data-driven model	MATLAB 9.5 (R 2018b) Classification Learner ann toolbox
Chapter 6: How Can Process Safety and a Risk Management Approach Guide Pandemic Risk Management?	*To devise semi-mechanistic models for assessing pandemic risk*	• Integrating data-based and science-based models by capturing temporal changes in the parameter of the SEIQRD model using ANN • Handling low-quality data by assigning a higher weight to the more reliable fatality data	MATLAB 9.5 (R 2018b) ode45 patternsearch ann toolbox

	To establish similirities between process safety principles and epidemiology	• Capturing uncertainty in the parameters using Monte Carlo simulation • Layer of protection analysis of the pandemic risk management	
Chapter 7: Pandemic Risk Assessment and Management in Bayesian Framework	*To devise semi-mechanistic models for assessing pandemic risk*	• Integrating data-based and science-based models • Treating uncertainty using the parameter sharing feature of the hierarchical Bayesian structure • Impact analysis for random parameters using relevance-based reasoning of Bayesian framework	PyMC3 3.10 pm.Model pm.sample arviz pyplot pandas GeNIe 2.2
Chapter 8: Pandemic Risk Management using Engineering Safety Principles	*To establish similarities between process safety principles and epidemiology*	• precautionary and ALARP-based risk management • risk-reducing strategies into hierarchical safety frameworks • developing event tree model of pandemic risk management for distinct risk-reducing strategies	MATLAB 9.5 (R 2018b) ode45 patternsearch

List of symbols and abbreviations

Symbols	Meanings
ALARP	as low as reasonably practicable
ANN	artificial neural network
ARIMA	autoregressive integrated moving average
COVID-19	Coronavirus disease of 2019
ETS	exponential smoothing
LOPA	layer of protection analysis
LSTM	long short-term memory
PDNN	process dynamics-guided neural network
QbT	quality-based training
SEIQRD	susceptible, exposed, infected, quarantined, recovered, deceased
HBN-SEIQRD	hierarchical Bayesian network-based susceptible, exposed, infected, quarantined, recovered, deceased
TN	Trillion
V-MFO	variable mosquito flying optimization

References

Alauddin, M., Khan, F., Imtiaz, S., & Ahmed, S. (2018). A Bibliometric Review and Analysis of Data-Driven Fault Detection and Diagnosis Methods for Process Systems. *Industrial and Engineering Chemistry Research*, *57*(32), 10719–10735. https://doi.org/10.1021/acs.iecr.8b00936

Camp, K. M., Mead, D., Reed, S. B., Sitter, C., & Wasilewski, D. (2020). From the barrel to the pump: the impact of the COVID-19 pandemic on prices for petroleum products. *Monthly Labor Review, US Bureau of Labor Statistics, October, 1-14*. https://doi.org/10.21916/mlr.2020.24

OECD (2020). The impact of Coronavirus (COVID-19) and the global oil price shock on the fiscal position of oil-exporting developing countries. *OECD Policy Responses to Coronavirus (COVID-19)*, OECD Publishing, Paris. https://doi.org/10.1787/8bafbd95-en.

Fisher, O.J., Watson, N.J., Escrig, J.E., Witt, R., Porcu, L., Bacon, D., Rigley, M., Gomes, R.L. (2020). Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. *Computers & Chemical Engineering.* 140, 106881. https://doi.org/10.1016/j.compchemeng.2020.106881

Khan, F. I., & Abbasi, S. A. (1999). Major accidents in process industries and an analysis of causes

and consequences. *Journal of Loss Prevention in the Process Industries*, *12*(5), 361–378. https://doi.org/10.1016/S0950-4230(98)00062-X

Mannan, M. S., Reyes-Valdes, O., Jain, P., Tamim, N., & Ahammad, M. (2016). The Evolution of Process Safety: Current Status and Future Direction. *Annual Review of Chemical and Biomolecular Engineering*, *7*, 135–162. https://doi.org/10.1146/annurev-chembioeng-080615-033640

Niu, G., Yang, B. S., & Pecht, M. (2010). Development of an optimized condition-based maintenance system by data fusion and reliability-centered maintenance. *Reliability Engineering and System Safety*, *95*(7), 786–796. https://doi.org/10.1016/j.ress.2010.02.016

Pawar, S., San, O., Aksoylu, B., Rasheed, A., & Kvamsdal, T. (2021). Physics guided machine learning using simplified theories. *Physics of Fluids*, *33*(1), 011701. https://doi.org/10.1063/5.0038929

Ren, X., Zhu, K., Cai, T., & Li, S. (2017). Fault Detection and Diagnosis for Nonlinear and Non-Gaussian Processes Based on Copula Subspace Division. *Industrial and Engineering Chemistry Research*, *56*(40), 11545–11564. https://doi.org/10.1021/acs.iecr.7b02419

Svendsen, D. H., Piles, M., Munoz-Mari, J., Luengo, D., Martino, L., & Camps-Valls, G. (2021). Integrating Domain Knowledge in Data-Driven Earth Observation With Process Convolutions. *IEEE Transactions on Geoscience and Remote Sensing*, 1–15. https://doi.org/10.1109/TGRS.2021.3059550

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis part I: Quantitative model-based methods. *Computers and Chemical Engineering*, *27*(3), 293–311. https://doi.org/10.1016/S0098-1354(02)00160-6

World Economic Forum, W., & Accenture. (2017). Digital Transformation Initiative: Oil and Gas Industry. *World Economic Forum*.

Chapter 2

Literature Survey

This chapter presents a survey of literature on data-driven and semi-mechanistic models for process safety and pandemic risk management. The literature on process fault detection and diagnosis with a focus on handling data quality and integrating mechanistic knowledge in data-driven models is reviewed. Similarly, the literature related to pandemic risk management is also divided in two sections: model development and parameter estimation.

2.1 Data-driven fault detection and diagnostic techniques

Many techniques of process fault detection and diagnosis (FDD) have been developed over the last four decades. They can broadly be divided in two categories: model-based and data-based. Model-based or mechanistic or first principle methods are based on rigorous mathematical formulations for representing conservation laws, domain knowledge, physical principles, and phenomenological behaviors (Liu, McDermid, & Chen, 2010; Venkatasubramanian, Rengaswamy, Kavuri, & Yin, 2003). These methods have been extensively used in process systems owing to their robustness and reliability. However, they are not easily implementable for early fault detection of complex processes. On the other hand, data-driven methods are based on process measurements, which do not require a priori quantitative or qualitative process knowledge (Isermann, 2005; Staroswiecki, 2000). Alauddin, Khan, Imtiaz, and Ahmed (2018) classified evolution and development of research on FDD in four categories namely, formulation of basic data-driven algorithms, advancement of the algorithms, applications of data-based algorithms in process systems, and development and application of advanced hybrid techniques in process systems. Some of the recent studies on survey of data-driven FDD includes: (Alauddin et al., 2018; Arunthavanathan, Khan, Ahmed, & Imtiaz, 2021; Dai & Gao, 2013; Gao, Cecati, & Ding, 2015;

Ge, 2017; Qin, 2012; Tidriri, Chatti, Verron, & Tiplica, 2016; Yin, Ding, Haghani, Hao, & Zhang, 2012; Yin, Ding, Xie, & Luo, 2014). The following section describes commonly used data-based models of fault detection and diagnosis of process systems.

Principal component analysis and partial least squares-based FDD methods: Principal component analysis (PCA) and partial least squares (PLS) are two leading multivariate fault detection and diagnostic approaches in the process systems (Alauddin et al., 2018). These methods are based on data projection to a lower-dimensional space for achieving fault detection and diagnosis (Kourti & MacGregor, 1995). The PCA and PLS-based methods can handle large numbers of highly correlated data, measurement errors, and missing data. The dimension reduction feature of these algorithms makes them ideal candidates in combining with other algorithms for handling big data. The major drawback of PCA and PLS based nethods is that they are not optimal for non-Gaussian and nonlinear data (Lawrence, 2005). Besides, Pearson's dependence on the covariance matrix makes PCA sensitive to outliers (Yu, Khan, & Garaniya, 2016). PLS does not perform well when input space has significant variation orthogonal to the output space. Thus numerous advanced data-based FDD algorithms have been developed over the last four decades such as, kernel principal component analysis (KPCA), Multiway PCA, multi-block PCA, hierarchical PCA, recursive PCA, sparse PCA, and moving window KPCA. Similarly, some of the variants of PLS are as follows: orthogonal projections to latent structures (O-PLS), which removes variation from descriptor variables that are not correlated to property variables, total projection to latent structure (TPLS) which makes further decomposition on particular subspaces, multi-way partial least squares (MPLS), dynamic PLS and recursive PLS. Kaspar and Harmon Ray (1993) proposed dynamic PLS for capturing the time-varying nature of process data. Qin (1998) formulated recursive partial

least squares (RPLS) algorithms for online process modeling to adapt process changes and offline modeling to deal with huge data samples.

Independent component analysis-based FDD methods: Independent component analysis (ICA) was developed to handle non-Gaussian data that are frequently encountered in process systems. It is based on higher-order statistics (HOS), such as kurtosis and negative entropy. Unlike PCA, ICA vectors are not orthogonal, and each component is equally important. Some of the key ICA algorithms include FastICA, simultaneous third-order tensor diagonalization (STOTD), joint approximate diagonalization of Eigen matrices (JADE), higher-order eigenvalue decomposition (HOEVD), maximal diagonality (MD), and maximum likelihood (ML) (De Lathauwer, De Moor, & Vandewalle, 2000). The limitation of the conventional approach of ICA is that the independent components cannot be ranked to the amount of variance explained by them (Chawla, 2011). Also, ICA is limited to linear systems and uses kernel-based mapping to deal with nonlinear systems that become computationally expensive with large samples. Moreover, fault identification and diagnosis are difficult due to the irreversible nature of kernel mapping (Deng, Tian, & Chen, 2013). Many researchers came up with modifications in the existing algorithms to deal with these limitations. For instance, Zhang and Zhang (2010) introduced modified ICA based on particle swarm optimization (PSO). Stefatos and Hamza (2010) used dynamic independent component analysis (DICA) methods to capture dynamic patterns. Zhang and Qin (2007) proposed a multiway kernel ICA (MKICA) method for dealing with nonlinearities in industrial systems. Rashid and Yu (2012) introduced the concept of hidden Markov's model in ICA for handling multimodal problems of industrial processes. Cai and Tian (2014) and Cai, Tian and Chen (2014) proposed robust ICA for handling outliers and noisy data. Tong, Palazoglu and Yan (2014) suggested improved ICA based on ensemble learning and Bayesian inference. Du et al (2017) introduced

fault-related kernel independent component analysis (FKICA), which decomposes data into four subspaces and makes the algorithm more sensitive to specific faults.

Copula-based fault detection and diagnosis methods: Vine copula-based methods are other potential tools for handling non-Gaussian data. Unlike PCA and PLS which are based on dimensionality reduction, the vine copula-based FDD tools exploit features correlation to estimate copula density and joint probability distribution for reckoning the fault probability. One of the C-vine-based method's intrinsic limitations is its inability to maximize the correlation structure. Cui and Li (2020) used a combination of C-vine and D-vine copula to capture both the stronger and weaker correlations by harnessing the maximum information coefficient (MIC) (Reshef, Reshef, Sabeti & Mitzenmacher, 2014). Ren et al. (2017) proposed a copula subspace division-based fault diagnosis technique.

Gaussian mixture model of fault detection and diagnosis: Mixture models are the other approach explored to deal with non-Gaussian data. The Gaussian mixture models (GMMs) are statistical methods based on the weighted sum of probability density functions of multiple Gaussian distributions. The number of Gaussian components is approximated by the expectation-maximization (EM) algorithm (Yu, 2013; Yu & Qin, 2009). The E-step determines the posterior probability, given all the parameters and the prior probabilities, while M-step finds all the parameters. The maximum likelihood solution is obtained by repeating the E- and M-steps iteratively. Thissen et al (2005) came up with GMM based on Bayesian information criteria (BIC). Li, Qin, and Yuan (2016) devised a framework of root cause analysis for locating faults for stationary and dynamic processes. The authors proposed a dynamic time warping-based causality analysis for digging the causality relations of the candidates. Yu (2012a) devised a nonlinear kernel Gaussian mixture model (NKGMM) for handling nonlinearities. Yu and Qin (2008) proposed a

finite Gaussian mixture model (FGMM) based on the Figueiredo–Jain (F–J) algorithm for optimizing Gaussian components and estimating statistical distributions. This formalism resulted in higher detection rates, lower false alarm rates, and shorter delays in fault detection. Yu, Khan and Garaniya (2015a) proposed a probabilistic multivariate method for fault diagnosis of industrial processes. The study employed a Gaussian copula based on rank correlation to model dependencies and nonlinearity of process variables. The technique is useful in handling nonlinearities; however, it requires a longer computational time for convergence. Similarly, Yu, Khan and Garaniya (2015b) devised a three-layered nonlinear Gaussian belief network (NLGBN) for fault diagnosis in industrial systems by introducing G-index, a novel parameter for process monitoring based on expectation maximization.

Bayesian network and hidden Markov models of FDD: Bayesian network (BN) and hidden Markov models (HMM) are tools for handling uncertainty by using probabilistic reasoning. The HMM is an efficient tool to estimate probability distributions of state transitions and unobservable states of a process. Smyth (1994) employed the HMM for continuous monitoring of complex dynamic systems. Yu (2012b) proposed the multiway discrete hidden Markov model (MDHMM) for fault detection and classification for handling uncertainty. Bayesian networks (BNs) are a commonly used technique for fault diagnosis. A BN can also be used to model hierarchical levels of multiple causes and effects. Verron et al. (2006) proposed a BN-based classifier for chemical process fault diagnosis. A dynamic Bayesian network (DBN) for process monitoring was implemented by (Yu & Rashid, 2013). Zhang and Dong (2014) proposed a multiple time-slice DBN-based FDD for handling dynamic nature and missing data. Amin et al. (2019b) proposed a dynamic Bayesian anomaly index (DBAI)-based thresholding technique for fault detection and diagnosis.

Machine learning-based models for fault detection and diagnosis: Many machine learning techniques have been successfully tested for fault detection and diagnostic methods of process systems. The artificial neural network (ANN) (Agatonovic-Kustrin & Beresford, 2000; Schmidhuber, 2015), support vector machines (SVM) (Burges, 1998; Cortes & Vapnik, 1995; Widodo & Yang, 2007), case-based reasoning (CBR) (Zhao et al., 2017), classification and regression tree (CART) (Lawrence & Wright, 2001) are some of the prominent methods in this category. The applications of ML algorithms in chemical health and safety studies can date back to the mid-1990s. However, the rapid advancement of computational resources and sophisticated algorithms over the last decade accelerated the neural network domain's research. Artificial neural network (ANN) has been inspired by biological neural systems. A standard ANN consists of many processors, also called neurons, which generate a sequence of real-valued activations (Agatonovic-Kustrin & Beresford, 2000; Schmidhuber, 2015). Recurrent neural networks (RNNs) and their variants have been developed to deal with complex sequential problems of time series data. Long short-term memory (LSTM) has been employed for fault detection and diagnosis for several processing applications (Li et al., 2017; Zhao, Sun, & Jin, 2018). Bidirectional recurrent neural networks (Zhang, Bi, & Qiu, 2020) and convolutional neural networks (Fuan, Hongkai, Haidong, Wenjing, & Shuaipeng, 2017; Wu & Zhao, 2018) have also been successfully exploited for the fault diagnosis of chemical processes. A comprehensive review of the application of machine learning and deep learning methods have in chemical safety is presented in (Arunthavanathan et al., 2021; Jiao, Hu, Xu, & Wang, 2020).

Hybrid data-driven models for fault detection and diagnosis: Many hybrid methods have been formulated for efficient process monitoring and FDD by exploiting the diagnosis capability of Bayesian. Alauddin, Khan, Imtiaz, and Ahmed (2018) illustrated how Bayesian has evolved as

the choice of recombinant and hybrid algorithm for detection and diagnosis of process faults. Mallick and Imtiaz (2013) combined PCA with BN for improving diagnostic performance. Gharahbagheri et al. (Gharahbagheri, Imtiaz, & Khan, 2017) proposed a data-driven BN learning algorithm using Granger causality and transfer entropy. Amin, Imtiaz, and Khan (2018) presented a hybrid approach to integrate the PCA and T^2 statistics with a Bayesian network (BN) model. A combination of the hidden Markov model (HMM) and a BN was put forward by (Galagedarage Don & Khan, 2019) for detection and diagnosis of process abnormalities.

Numerous hybrid algorithms have been asserted to exhilarate the training of artificial neural networks. Barakat, Druaux, Lefebvre, Khalil, and Mustapha (2011) developed a self-adaptive FDD based on artificial neural networks combined with an advanced signal processing method. Nawi, Khan, and Rehman (2013) combined Cuckoo Search (CS) and Levenberg Marquardt algorithm to train neural networks. Dehuri and Cho (2009) proposed a multi-objective pareto-based particle swarm method for optimizing architectural complexity and classification accuracy of a neural network. Subudhi and Jena (2011a) employed opposition-based differential evolution (ODE) for training a feed-forward neural network. Juang (2004) trained a recurrent neural network with a hybrid GA-PSO algorithm using a PSO with a master-slave configuration. Subudhi and Jena (2011b) proposed a memetic algorithm-based differential evolution back-propagation (DEBP) for training a multilayer perceptron by exploiting local and global search spaces.

2.1.1 Data-driven models for handling low-quality data

Many robust techniques have been devised to deal with low-quality data. The missing data problem is handled by mean substitution, single regression imputation, last observation carried forward (LOCF), and modern methods based on maximum likelihood and Bayesian multiple imputations (Imtiaz & Shah, 2008; Jin, Wang, Huang, & Forbes, 2012; Khatibisepehr & Huang, 2008; Schafer

& Graham, 2002). Bayesian network (BN) and hidden Markov models (HMM) are efficient tools based on probabilistic reasoning for handling uncertainty. The HMM estimates the probability distributions of state transitions and unobservable states of the process. Independent component analysis (ICA) was developed to handle the non-Gaussian data that are frequently encountered in process systems. Rashid and Yu (2012) introduced the concept of hidden Markov's model in ICA for handling multimodal problems of industrial processes. Cai and Tian (2014) and Cai, Tian, and Chen (2014) proposed robust ICA for handling outliers and noisy data. Vine copula-based methods are other potential tools for handling non-Gaussian data.

Robust statistics such as the S-estimators (Davies, 2007), M-estimators (Maronna, 2007), and minimum covariance determinant (MCD) estimators (Baz-Lomba, Harman, Reid, & Thomas, 2017; Seheult, Green, Rousseeuw, & Leroy, 1989) have been developed for detecting outliers. Zhu, Shi, Song, Tao, and Tan (2020) proposed an information concentrated variational auto-encoder (IFCVAE) by dividing latent variables into quality-related and unrelated spaces. The other weight-based studies include Spherical PCA (sPCA) (Stanimirova, Daszykowski, & Walczak, 2007), a weighted PPCA (WPPCA) (Yuan et al., 2017), divergence-based robust ICA (Chen, Hung, Komori, Huang, & Eguchi, 2013; Mihoko & Eguchi, 2002), a meta gradient descent-based reweighting (Ren, Zeng, & Yang, 2018) and robust Bayesian method (Chen & Ge, 2020).

2.1.2 Survey on integrating mechanistic and data-driven models

The generalization characteristics of data-driven methods have been improved by integrating mechanistic models. Table 2.1 presents hybrid formalism for integrating process knowledge in a neural network framework. Psichogios and Ungar (1992) employed a hybrid model in which the neural network component process parameters of the first principles model. Similarly, Thompson and Kramer (1994) employed a hybrid model based on a parametric model (that compensates for

sparse data) and a neural network that accounts for uncertainty and bias. Xiong and Jutan (2002) investigated a hybrid model-based control strategy using a parallel structure where a neural network was used to compensate for the mismatch of a detailed and approximate mechanistic model. Georgieva et al. (2003) employed a hybrid model in an industrial fed-batch evaporative crystallization process for predicting size distribution in a refining process. Mass and energy balances were captured in the mechanistic sub-model, the data-driven sub-model described growth rate, nucleation, and agglomeration parameters. Oliveira (2004) proposed a hybrid formalism with parallel and fused multiplication functions. Stewart and Ermon (2017) formulated a new approach to supervising neural networks by specifying mechanistic constraints in the output space. Azarpour et al. (2017) developed a generic framework based on the first principle model and an artificial neural network to study catalyst deactivation of fixed-bed catalytic reactors (FBCRs). Muralidhar et al. (2018) illustrated how incorporating domain constraints into the loss function can be used to ameliorate the modeling of a sparse and noisy dataset. Hendriks et al. (2020) presented a neural network-based model that explicitly satisfies known linear operator constraints. Wu et al. (2020) proposed three distinct hybrid formalisms based on a recurrent neural network to integrate physical knowledge in data-driven models.

Many studies conferred that meta-learning can substantially enhance the performance of deep neural networks. Van Lith et al. (2003) augmented a simple physical framework with fuzzy logic. Lee and Kang (2007) formulated a modified back-propagation neural network (BPN) based meta-model that ensures constraint feasibility of the approximate optimal solution. Alauddin et al. (2020) presented an ANN-based fault detection model that was tuned using a variable mosquito flying optimization (V-MFO) technique for maximizing fault detection rate (FDR) and minimizing false alarm rate (FAR).

Table 2.1: Hybrid formalisms for integrating process knowledge in neural network models

	Study	Hybrid Formalism	Case study
	(Psichogios & Ungar, 1992)	Process parameters of the first principle model were determined by neural network	Fed-batch bioreactor
	(Thompson & Kramer, 1994)	Parametric model that compensates for sparse data and neural network accounting for uncertainty and bias	Penicillin fermentation
	(Xiong & Jutan, 2002)	ANN was used to compensate for model mismatch of detailed and approximate mechanistic models.	Batch reactor; continuous stirred tank reactor
	(Bollas et al., 2003)	Neural model was used to refine the plant model prediction	Fluid catalytic cracking
	(Georgieva, Meireles, & Feyo de Azevedo, 2003)	Mass and energy balances were captured in the mechanistic sub-model the data-driven sub-model described the growth rate, nucleation and agglomeration parameters.	Fed-batch evaporative crystallization process
First principle models	(Oliveira, 2004)	Parallel and fused mechanistic part and a data-driven part.	Recombinant protein and baker's yeast production
	(Safari, Shabani, & Simon, 2014)	Neural network fused the outputs of multiple Kalman filters	Industrial sensor fusion
	(Yang, Dai, Tang, Xuan, & Cao, 2020)	Integrating deep neural network with a physical lumped kinetic model	Automated FCC process.
	(Chen & Ierapetritou, 2020)	Serial, parallel, and combined structures of hybrid models.	Simulated reactor mode
	(Wu, Rincon, & Christofides, 2020)	Physics-based recurrent neural network (RNN)	Model predictive control
	(Ghosh, Moreira, & Mhaskar, 2021)	Data-driven model learns the residuals from the mechanistic model; mechanistic nonlinear model approximated by surrogate linear model and learning of residual using a data-driven model	Batch crystallization
Meta-modeling	(Tan & Li, 2002)	Mechanistic sub-model based on momentum balance and the data-driven model using Padé approximation	Hydraulic nonlinear system
	(Van Lith et al., 2003)	Simple physical framework augmented with fuzzy logic. Information about the	Experimental batch distillation column,

	dynamic behavior is incorporated in the form of prior knowledge	
(Lee & Kang, 2007)	Modified back-propagation neural network (BPN) based meta-model that ensures constraint feasibility of the approximate optimal solution.	Standard structural problems
(Alauddin, Khan, et al., 2020)	ANN-based fault detection model was tuned using a variable mosquito flying optimization technique for maximizing fault detection rate and minimizing false alarm rate.	Tennessee Eastman process

2.2 *Literature survey of epidemiological models*

Numerous mathematical methods and computational techniques based on differential equations, stochastic processes, statistical analysis, graph theory, artificial society, computer simulation, and geographic information systems have been developed to study epidemic disease transmissions. These can broadly be classified in two categories: statistical and mechanistic. Statistical approaches exploit data correlations to learn a functional dependence for predictions, while mechanistic models are based on physical laws such as population and/or transmission dynamics. Mechanistic models comprise compartmental, spatial, meta-population, network-based, and agent-based methods. In this section, we present an overview of compartmental models for the dynamics and control of infectious diseases.

2.2.1 *Compartmental epidemiological models*

Compartmental models are based on systems of ordinary differential equations that focus on the dynamic progression of a population through different epidemiological states (Chowell, 2017). The population is divided into distinct compartments, each having the same state of the epidemic. The SIR (susceptible, infected, recovered) and SEIR (susceptible, exposed, infected, recovered) are fundamental compartmental models. The SIR model presumes that the infected hosts become

contagious immediately after exposure to an infected carrier, whereas the SEIR model considers the latency between exposures and infectious periods (Anderson & May, 1979; Hethcote, 1976; Hiorns & MacDonald, 1982; Kermack & McKendrick, 1927). Many extended compartmental models have been developed to take into account isolation, quarantine, and hospitalization (Alauddin et al., 2020; Arik et al., 2020; Giordano et al., 2020; Hu et al., 2020; Ivorra, Ferrández, Vela-Pérez, & Ramos, 2020; Legrand, Grais, Boelle, Valleron, & Flahault, 2007; Lin et al., 2020; Paiva, Afonso, de Oliveira, & Garcia, 2020; Subramanian, He, & Pascual, 2021).

Many models have been developed for taking into account non-pharmaceutical interventions (NPIs) of isolation and quarantine of the exposed cases. Carvalho, da Silva and Charret (2019) presented a comparative analysis of mechanical and chemical control methods in restraining a pandemic. The modeling of comprehensive interventions, treatment, and other control actions can be found in several studies (Alam, Kabir, & Tanimoto, 2020; Browne, Gulbudak, & Webb, 2015; Chowell & Kiskowski, 2016; Colizza, Barrat, Barthelemy, Valleron, & Vespignani, 2007; De Visscher, 2020; Fast, Mekaru, Brownstein, Postlethwaite, & Markuzon, 2015; Jung, Lee, & Chowell, 2016; Lee, Lye, & Wilder-Smith, 2009; Rachah & Torres, 2016; Rizzo & Atti, 2008; Shen, Xiao, & Rong, 2015).

2.2.2 Parameter estimation in epidemiological models

The parameters of mechanistic models can be estimated by least-squares fitting (Gan, Tan, Mo, Li, & Huang, 2020; Kim, Ko, Kim, & Jung, 2020; Romero-Severson, Ribeiro, & Castro, 2018), maximum likelihood estimation (Bretó, 2018; Chowell, Nishiura, & Bettencourt, 2007; Kao & Eisenberg, 2018; White & Pagano, 2008; Wu & Riley, 2016), and approximate Bayesian computation (Almutiry & Deardon, 2020; Brown, Porter, Oleson, & Hinman, 2018; Chandra, 2020; Kypraios, Neal, & Prangle, 2017; McKinley et al., 2018; Minter & Retkute, 2019; Neal,

2019; Neal & Terry Huang, 2015; Price, Bean, Ross, & Tuke, 2016). Epidemiological modeling has been extensively studied using deterministic approaches (Carcione et al., 2020; Clancy & Piunovskiy, 2005; House & Keeling, 2008; Kumar et al., 2020; Lange, 2016; Lipsitch et al., 2003; Okyere et al., 2020; Otoo et al., 2020; Pollicott et al., 2012; Rivers et al., 2014; Sharkey, 2011; Zhou et al., 2014) and stochastic approaches (Alharthi, Kypraios, & O'Neill, 2019; Allen, 2017; Birrell, de Angelis, & Presanis, 2018; Bjørnstad, Finkenstädt, & Grenfell, 2002; Chao, Halloran, Obenchain, & Longini, 2010; Engbert, Rabe, Kliegl, & Reich, 2021; Fintzi, Cui, Wakefield, & Minin, 2017; Khan, Hussain, Zahri, Zaman, & Wannasingha Humphries, 2020; Kypraios et al., 2017; Lekone & Finkenstädt, 2006; O'Neill, 2002; Ponciano & Capistrán, 2011; Shangguan, Liu, Wang, & Tan, 2021; Taylor, Dushoff, & Weitz, 2016; Wang, Ji, Bi, & Liu, 2020). Pollicott et al. (2012) reconstructed a time-dependent transmission rate for the SIR model. Camacho et al. (2015) modeled the time-varying transmission parameter of SEIR by the Wiener process (also known as standard Brownian motion). Cauchemez et al. (2008) used a stochastic framework and Markov chain Monte Carlo (MCMC) to recover time-dependent transmission rate and other model parameters. Alauddin et al. (2020) proposed a semi-mechanistic artificial neural network-based SEIQRD model for capturing time varying parameters. Google Cloud researchers integrated the time series machine learning approach with a compartmental model to develop Google Cloud AI (Arik et al., 2020). Population genetic inference using coalescent models has been exploited in the parameter estimation of compartmental models (Dearlove & Wilson, 2013; Popinga, Vaughan, Stadler, & Drummond, 2014). Many studies have investigated branching processes to estimate basic reproduction number (R0) and the final size of the epidemic (Allen, 2015; Ball & Donnelly, 1993; Tritch & Allen, 2018).

Evolutionary computing approaches have also been explored for optimal parameter estimation for deterministic and stochastic epidemiological models. For instance, a genetic algorithm was employed for the parameter estimation of several epidemic diseases such as cholera (Akman & Schaefer, 2015), malaria (Davis et al., 2019), SARS (Isa Irawan & Amiroch, 2015; Yan & Zou, 2008), HIV-AIDS (Pedroso-Rodriguez, Marrero, & De Arazoza, 2003), and COVID-19 (Anđelić, Šegota, Lorencin, Mrzljak, & Car, 2021; Ding et al., 2021; Hosseini, Ghafoor, Sadiq, Guizani, & Emrouznejad, 2020; Monteiro, Gandini, & Schimit, 2020). Eastman et al. (2021) used differential evolution to fit their model to the epidemiological data of COVID-19. Similarly, swarm intelligence has been exploited to fitting epidemic models (Akman, Akman, & Schaefer, 2018; He, Peng, & Sun, 2020; Hosseini et al., 2020).

The predictive ability of a model can be improved by reducing variability in the forecasts. Xu et al. (2016) proposed a Bayesian non-parametric method for stochastic epidemic models using a Gaussian process. Oliveira et al. (2020) employed a Bayesian approach to the SIR model with correction for under-reporting of COVID-19 cases (Manevski, Ružić Gorenjec, Kejžar, & Blagus, 2020). Polo et al. (2020) used a Bayesian model for studying the spatio-temporal effect on health and fatality due to COVID-19. Bayesian inference was also used to construe the time dependence of the effective growth rate of newly infected cases (Das & Tiwari, 2021; Dehning et al., 2020). Farah et al. (2014) developed a Bayesian framework to estimate parameters of a computationally expensive dynamic epidemic model using time series epidemic data. Dehning et al. (2020) applied Bayesian inference based on Monte Carlo sampling to characterize the change points realized due to distinct mitigative measures. Smirnova et al. (2019) introduced the reconstruction of nonparametric time-dependent transmission rates by projecting onto a finite subspace spanned by Legendre polynomials for precise approximation of model parameters.

Numerous methods in the Bayesian framework were devised for handling data inconsistency and under-reported cases (Deardon et al., 2010; Gibson, Reich, & Sheldon, 2020; Russell et al., 2020; Sharmin, Glass, Viennet, & Harley, 2018; Taghizadeh, Karimi, & Heitzinger, 2020). Li et al. (2018) compared distinct models of varying complexity on three MCMC platforms: JAGS (Plummer, 2003), NIMBLE (Lawson, 2020), and Stan (Carpenter et al., 2017). Lytras et al.(2019) introduced FluHMM based on a Hidden Markov Model fitted in a Bayesian framework. Azmon et al. (2014) investigated the impact of delays and under-reporting by using a Bayesian semiparametric approach with penalized splines. Markov switching model has been employed to determine the epidemic and non-epidemic periods of surveillance data (Amorós, Conesa, López-Quílez, & Martinez-Beneito, 2020; Martínez-Beneito, Conesa, López-Quílez, & López-Maside, 2008). Dureau et al.(2013) employed an adjusted adaptive particle Markov chain Monte Carlo algorithm for handling parameter uncertainty. Bayesian inference using a Markov chain Monte Carlo (MCMC) approach to parameter estimation in epidemic models has been reported in several studies (Almutiry & Deardon, 2020; Brown et al., 2018; Cotta, Naveira-Cotta, & Magal, 2020; Deardon et al., 2010; Li et al., 2018; Marwa, Mwalili, & Mbalawata, 2019; Osei, Duker, & Stein, 2012; Taghizadeh et al., 2020; Touloupou, Finkenstädt, & Spencer, 2020). Lee and Mallick (2020) employed a hierarchical Bayesian approach for estimating COVID-19 spread curves by borrowing information from global data. However, the epidemic disease trajectory depends on demographics, economic status, degree of compliance of the population and societal impact, comorbidities, overall risk environment, and country vulnerability to biological threats.

2.3 Identified knowledge gaps and contributions of this work

A plethora of literature is available on fault detection and diagnosis of process systems using data-driven methodologies. Many of those promising tools have been successfully tested for the

industrial scale. This thesis presents several formalisms for improving data-driven and mechanistic models of process safety and pandemic risk management as discussed in Section 1.1-1.3.

This section summarizes thesis contribution to establish synergy between process safety and pandemic risk management. The methodologies to prevent, control, and mitigate infection are analogous to the hazard control and safety frameworks used in the process industries. Only a few studies outlined the process safety concepts for pandemic risk management. Brown, Amyotte and Vanberkel (2021) classified measures of restraining an epidemic disease into hierarchical process safety principles. Lindhout and Reniers (2020) proposed an integrated pandemics barrier model based on the sequential steps of an outbreak. However, synergy between the pandemic risk assessment and process safety has yet to be investigated. This thesis presents similarities between the SIR epidemiological model and the reaction kinetics model of a CSTR by highlighting resemblance in the conservation principles and factors governing contagion and reaction rates (Chapter 6). Also, several areas of similarities have been identified where process safety and epidemiology can benefit from each other. These include (i) early fault detection vs early case detection, (ii) identification of effective control mechanism, (iii) the fast response of public health vs operator response, (iv) effective resource allocation and mobilization, (v) identification of the most vulnerable elements, and (vi) application of expertise from similar outbreaks in the past vs use of historical process data (Chapter 6).

The other contributions of the thesis in this regard are as follows-

    a. *Pandemic risk management using precautionary principle:* Precautionary approach is a principle for making practical decisions under scientific uncertainty. Section 8.3.1 presents pandemic risk assessment using the precautionary approach based on the SEIQRD epidemiological model.

b. *ALARP approach to pandemic risk management:* The *ALARP (as low as reasonably practicable)* approach is based on risk-informed and cautionary thinking. The enforcement of risk-reducing measures, including no measure, has been studied using the ALARP approach to risk assessment. The quantitative risk estimated using the SEIQRD model and uncertainty in the parameters has been captured by using the Monte Carlo simulation (Section 8.3.3).

c. *Development of event tree model of pandemic risk management:* Event tree is a causal analytical technique used to identify potential accident scenarios and sequences in a process system. It can help to prevent negative outcomes using distinct safeguards taking into account whether installed safety barriers are functioning or not. The risk of the accident in process systems is minimized by activating safeguards. The safe working of these safeguards determines the intended risk in case of an accident. Similarly, the impact of a pandemic outbreak is largely governed by distinct measures taken to restrict, control and mitigate the outbreak. The thesis presents an event tree model of pandemic risk management using distinct risk-reducing strategies realized due to natural evolution, government interventions, societal responses, and individual practices (Section 8.3).

d. *The layer of protection analysis of pandemic risk management:* The layer of protection analysis (LOPA) is a semi-quantitative approach to risk assessment. The safety of chemical processing operations is ensured by using multiple layers of safety barriers such as basic process control, alarms, operator interventions, safety instrumented system, relief devices, and physical containments (Dowell, 1999). Finally, the plant and emergency response services are brought into operation to diminish the risk by de-escalating the abnormal situation. These are analogous to the preventive, controlling, and mitigative

strategies of handling a pandemic. Hygienic practices (e.g. frequent hand washing, use of mask) and government regulations such as the closure of schools and non-essential business, limiting gathering sizes, enforcing lockdown, and vaccination are extremely effective as control and preventive strategies for limiting risk in the pandemic model. The fatality risk is minimized by ensuring sophisticated treatment by extending healthcare facilities. This is similar to the plant and emergency response services of a process safety system. The thesis presents a layer of protection analysis for pandemic risk management using preventive, controlling, and mitigative strategies (Section 6.4).

List of symbols and abbreviations

Symbols	Meanings
ANN	artificial neural network
ARIMA	autoregressive integrated moving average
ARMA	autoregressive moving average
DGMM	dynamic Gaussian mixture model
DICA	dynamic independent component analysis
EKF	extended Kalman filter
EM	expectation maximization
ES	expert System
EWMA	exponentially weighted moving average
FDD	fault detection and diagnosis
FGMM	finite Gaussian mixture model
FKICA	fault-related kernel independent component analysis
GBM	gray-box model
GMM	Gaussian mixture model
ICA	independent component analysis
JADE	joint approximate diagonalization of Eigen matrices
KPCA	kernel principal component analysis
LFMs	latent force models
LPP	locality preserving projections
MIPLS	multiway interval partial least squares
MKICA	multiway kernel ICA
MKPCA	multiway kernel principal component analysis
ML	maximum Likelihood
MPLS	multi-way partial least square
NKGMM	nonlinear kernel Gaussian mixture model
NLGBN	nonlinear Gaussian belief network
O-PLS	orthogonal projections to latent structures
PCA	principle component analysis
PLS	partial least square
QTA	qualitative trend analysis
RNN	recurrent neural networks
RPLS	recursive partial least squares
SDG	signed directed graph
SEIQRD	susceptible, exposed, infected, quarantined, recovered, deceased
SEIR	susceptible, exposed, infected, and recovered
SIR	susceptible, infected, recovered
SPC	statistical process control

sPCA	sparse principal component analysis
STOTD	simultaneous third-order tensor diagonalization
SVM	support vector machine
TPLS	total projection to latent structure
UKF	unscented Kalman filter

References

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Akman, D., Akman, O., & Schaefer, E. (2018). Parameter Estimation in Ordinary Differential Equations Modeling via Particle Swarm Optimization. *Journal of Applied Mathematics*, *2018*, 1–9. https://doi.org/10.1155/2018/9160793

Akman, O., & Schaefer, E. (2015). An evolutionary computing approach for parameter estimation investigation of a model for cholera. *Journal of Biological Dynamics*, *9*, 147–158. https://doi.org/10.1080/17513758.2015.1039608

Alam, M., Kabir, K. M. A., & Tanimoto, J. (2020). Based on mathematical epidemiology and evolutionary game theory, which is more effective: Quarantine or isolation policy. *Journal of Statistical Mechanics: Theory and Experiment*, *2020*(3), 033502. https://doi.org/10.1088/1742-5468/ab75ea

Alauddin, M., Islam Khan, M. A., Khan, F., Imtiaz, S., Ahmed, S., & Amyotte, P. (2020). How can process safety and a risk management approach guide pandemic risk management? *Journal of Loss Prevention in the Process Industries*, *68*, 104310. https://doi.org/10.1016/j.jlp.2020.104310

Alauddin, M., Khan, F., Imtiaz, S., & Ahmed, S. (2018). A Bibliometric Review and Analysis of Data-Driven Fault Detection and Diagnosis Methods for Process Systems. *Industrial and Engineering Chemistry Research*, *57*(32), 10719–10735. review-article. https://doi.org/10.1021/acs.iecr.8b00936

Alauddin, M., Khan, F., Imtiaz, S., & Ahmed, S. (2020). A variable mosquito flying optimization-based hybrid artificial neural network model for the alarm tuning of process fault detection systems. *Process Safety Progress*, *39*(S1), e12122. https://doi.org/10.1002/prs.12122

Alharthi, M., Kypraios, T., & O'Neill, P. D. (2019). Bayes factors for partially observed stochastic

epidemic models. *Bayesian Analysis*, *14*(3), 907–936. https://doi.org/10.1214/18-BA1134

Allen, L. J. S. (2015). Applications of Multi-Type Branching Processes. *Stochastic Population and Epidemic Models*, 21–27. https://doi.org/10.1007/978-3-319-21554-9_3

Allen, L. J. S. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, *2*(2), 128–142. https://doi.org/10.1016/j.idm.2017.03.001

Almutiry, W., & Deardon, R. (2020). Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease using Approximate Bayesian Computation. *International Journal of Biostatistics*, *16*(1), 20170092. https://doi.org/10.1515/ijb-2017-0092

Amin, M. T., Imtiaz, S., & Khan, F. (2018). Process system fault detection and diagnosis using a hybrid technique. *Chemical Engineering Science*, *189*, 191–211. https://doi.org/10.1016/j.ces.2018.05.045

Amin, M. T., Khan, F., & Imtiaz, S. (2019). Fault detection and pathway analysis using a dynamic Bayesian network. *Chemical Engineering Science*, *195*, 777–790. https://doi.org/10.1016/j.ces.2018.10.024

Amorós, R., Conesa, D., López-Quílez, A., & Martinez-Beneito, M. A. (2020). A spatio-temporal hierarchical Markov switching model for the early detection of influenza outbreaks. *Stochastic Environmental Research and Risk Assessment*, *34*(2), 275–292. https://doi.org/10.1007/s00477-020-01773-5

Anđelić, N., Šegota, S. B., Lorencin, I., Mrzljak, V., & Car, Z. (2021). Estimation of covid-19 epidemic curves using genetic programming algorithm. *Health Informatics Journal*, *27*(1), 1–40. https://doi.org/10.1177/1460458220976728

Anderson, R. M., & May, R. M. (1979). Population biology of infectious diseases: Part I. *Nature*, *280*(5721), 361–367. https://doi.org/10.1038/280361a0

Arik, S.O., Li, C.-L., Yoon, J., Sinha, R., Epshteyn, A., Le, L.T., Menon, V., Singh, S., Zhang, L., Yoder, N., Nikoltchev, M., Sonthalia, Y., Nakhost, H., Kanal, E., Pfister, T. (2020). Interpretable Sequence Learning for COVID-19 Forecasting. *ArXiv*. Retrieved from http://arxiv.org/abs/2008.00646

Arunthavanathan, R., Khan, F., Ahmed, S., & Imtiaz, S. (2021). An analysis of process fault diagnosis methods from safety perspectives. *Computers and Chemical Engineering*, *145*, 107197. https://doi.org/10.1016/j.compchemeng.2020.107197

Azarpour, A., N.G. Borhani, T., R. Wan Alwi, S., A. Manan, Z., & I. Abdul Mutalib, M. (2017). A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors. *Chemical Engineering Research and Design*, *117*, 149–167. https://doi.org/10.1016/j.cherd.2016.10.024

Azmon, A., Faes, C., & Hens, N. (2014). On the estimation of the reproduction number based on misreported epidemic data. *Statistics in Medicine*, *33*(7), 1176–1192. https://doi.org/10.1002/sim.6015

Ball, F., & Donnelly, P. (1993). Branching Process Approximation of Epidemic Models. *Theory of Probability & Its Applications*, *37*(1), 119–121. https://doi.org/10.1137/1137024

Barakat, M., Druaux, F., Lefebvre, D., Khalil, M., & Mustapha, O. (2011). Self adaptive growing neural network classifier for faults detection and diagnosis. *Neurocomputing*, *74*(18), 3865–3876. https://doi.org/10.1016/j.neucom.2011.08.001

Baz-Lomba, J. A., Harman, C., Reid, M., & Thomas, K. V. (2017). Passive sampling of wastewater as a tool for the long-term monitoring of community exposure: Illicit and prescription drug trends as a proof of concept. *Water Research*, *121*, 221–230. https://doi.org/10.1016/j.watres.2017.05.041

Birrell, P. J., de Angelis, D., & Presanis, A. M. (2018). Evidence synthesis for stochastic epidemic models. *Statistical Science*, *33*(1), 34–43. https://doi.org/10.1214/17-STS631

Bjørnstad, O. N., Finkenstädt, B. F., & Grenfell, B. T. (2002). Dynamics of measles epidemics: Estimating scaling of transmission rates using a Time series SIR model. *Ecological Monographs*, *72*(2), 169–184. https://doi.org/10.1890/0012-9615(2002)072[0169:DOMEES]2.0.CO;2

Bollas, G. M., Papadokonstadakis, S., Michalopoulos, J., Arampatzis, G., Lappas, A. A., Vasalos, I. A., & Lygeros, A. (2003). Using hybrid neural networks in scaling up an FCC model from a pilot plant to an industrial unit. *Chemical Engineering and Processing: Process Intensification*, *42*(8–9), 697–713. https://doi.org/10.1016/S0255-2701(02)00206-4

Bretó, C. (2018). Modeling and inference for infectious disease dynamics: A likelihood-based approach. *Statistical Science*, *33*(1), 57–69. https://doi.org/10.1214/17-STS636

Brown, K. R., Vanberkel, P., Khan, F., Amyotte, P., Application of bow tie analysis and inherently safer design to the novel coronavirus hazard (2021), *Process Safety and Environmental Protection*, 152, 701-718. https://doi.org/10.1016/j.psep.2021.06.046

Brown, G. D., Porter, A. T., Oleson, J. J., & Hinman, J. A. (2018). Approximate Bayesian computation for spatial SEIR(S) epidemic models. *Spatial and Spatio-Temporal Epidemiology*, *24*, 27–37. https://doi.org/10.1016/j.sste.2017.11.001

Browne, C., Gulbudak, H., & Webb, G. (2015). Modeling contact tracing in outbreaks with application to Ebola. *Journal of Theoretical Biology*, *384*, 33–49. https://doi.org/10.1016/j.jtbi.2015.08.004

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Cai, L., & Tian, X. (2014). A new fault detection method for non-Gaussian process based on robust independent component analysis. *Process Safety and Environmental Protection*, *92*(6), 645–658. https://doi.org/10.1016/j.psep.2013.11.003

Cai, L., Tian, X., & Chen, S. (2014). A process monitoring method based on noisy independent component analysis. *Neurocomputing*, *127*, 231–246. https://doi.org/10.1016/j.neucom.2013.07.029

Camacho, A., Kucharski, A., Aki-Sawyerr, Y., White, M.A., Flasche, S., Baguelin, M., Pollington, T., Carney, J.R., Glover, R., Smout, E., Tiffany, A., Edmunds, W.J., Funk, S. (2015). Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements: a Real-time Modelling Study. *PLoS Currents*, *Feb*. https://doi.org/10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2

Carcione, J. M., Santos, J. E., Bagaini, C., & Ba, J. (2020). A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model. *Frontiers in Public Health*, *8*, 230. https://doi.org/10.3389/fpubh.2020.00230

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–37. https://doi.org/10.18637/jss.v076.i01

Carvalho, S. A., da Silva, S. O., & Charret, I. da C. (2019). Mathematical modeling of dengue epidemic: control methods and vaccination strategies. *Theory in Biosciences*, *138*(2), 223–239. https://doi.org/10.1007/s12064-019-00273-7

Cauchemez, S., Valleron, A. J., Boëlle, P. Y., Flahault, A., & Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, *452*, 750–754. https://doi.org/10.1038/nature06732

Chandra, V. (2020). Stochastic Compartmental Modelling of SARS-CoV-2 with Approximate Bayesian Computation. *MedRxiv*. https://doi.org/10.1101/2020.03.29.20046862

Chao, D. L., Halloran, M. E., Obenchain, V. J., & Longini, I. M. (2010). FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology*, *6*(1), e1000656. https://doi.org/10.1371/journal.pcbi.1000656

Chawla, M. P. S. (2011). PCA and ICA processing methods for removal of artifacts and noise in electrocardiograms: A survey and comparison. *Applied Soft Computing Journal*, *11*(2), 2216–2226. https://doi.org/10.1016/j.asoc.2010.08.001

Chen, Y., & Ierapetritou, M. (2020). A framework of hybrid model development with identification of plant-model mismatch. *AIChE Journal*, *66*(10), 1–16. https://doi.org/10.1002/aic.16996

Chowell, G. (2017). Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*, *2*(3), 379–398. https://doi.org/10.1016/j.idm.2017.08.001

Chowell, G., & Kiskowski, M. (2016). Modeling ring-vaccination strategies to control ebola virus disease epidemics. In *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases* (pp. 71–87). https://doi.org/10.1007/978-3-319-40413-4_6

Chowell, G., Nishiura, H., & Bettencourt, L. M. A. (2007). Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface*, *4*(12), 154–166. https://doi.org/10.1098/rsif.2006.0161

Clancy, D., & Piunovskiy, A. B. (2005). An explicit optimal isolation policy for a deterministic epidemic model. *Applied Mathematics and Computation*, *163*(3), 1109–1121. https://doi.org/10.1016/j.amc.2004.06.028

Colizza, V., Barrat, A., Barthelemy, M., Valleron, A. J., & Vespignani, A. (2007). Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Medicine*, *4*(1), 0095–0110. https://doi.org/10.1371/journal.pmed.0040013

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20, 273-297*. https://doi.org/10.1023/A:1022627411411

Cotta, R. M., Naveira-Cotta, C. P., & Magal, P. (2020). Mathematical Parameters of the COVID-19 Epidemic in Brazil and Evaluation of the Impact of Different Public Health Measures. *Biology*, *9*(8), 220. https://doi.org/10.3390/biology9080220

Cui, Q., & Li, S. (2020). Process monitoring method based on correlation variable classification and vine copula. *Canadian Journal of Chemical Engineering*, *98*(6), 1411–1428. https://doi.org/10.1002/cjce.23702

Dai, X., & Gao, Z. (2013). From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis. *IEEE Transactions on Industrial Informatics*, *9*(4), 2226–2238. https://doi.org/10.1109/TII.2013.2243743

Das, S. S., & Tiwari, A. K. (2021). Understanding international and domestic travel intention of Indian travellers during COVID-19 using a Bayesian approach. *Tourism Recreation Research*, *46*(2), 228–244. https://doi.org/10.1080/02508281.2020.1830341

Davies, P. L. (2007). Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *The Annals of Statistics*, *15*(3), 1269–1292. https://doi.org/10.1214/aos/1176350505

Davis, J. K., Gebrehiwot, T., Worku, M., Awoke, W., Mihretie, A., Nekorchuk, D., & Wimberly, M. C. (2019). A genetic algorithm for identifying spatially-varying environmental drivers in a malaria time series model. *Environmental Modelling and Software*, *119*, 275–284. https://doi.org/10.1016/j.envsoft.2019.06.010

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). An introduction to independent component analysis. *Journal of Chemometrics*, *14*(3), 123–149. https://doi.org/10.1002/1099-128X(200005/06)14:3<123::AID-CEM589>3.0.CO;2-1

de Oliveira, A. C. S., Morita, L. H. M., da Silva, E. B., Zardo, L. A. R., Fontes, C. J. F., & Granzotto, D. C. T. (2020). Bayesian modeling of COVID-19 cases with a correction to account for under-reported cases. *Infectious Disease Modelling*, *5*, 699–713. https://doi.org/10.1016/j.idm.2020.09.005

De Visscher, A. (2020). The COVID-19 pandemic: model-based evaluation of non-pharmaceutical interventions and prognoses. *Nonlinear Dynamics*, *101*(3), 1871–1887. https://doi.org/10.1007/s11071-020-05861-7

Deardon, R., Brooks, S.P., Grenfell, B.T., Keeling, M.J., Tildesley, M.J., Savill, N.J., Shaw, D.J., Woolhouse, M.E.J. (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, *20*(1), 239–261.

Dearlove, B., & Wilson, D. J. (2013). Coalescent inference for infectious disease: meta-analysis of hepatitis C. *Philosophical Transactions of the Royal Society B: Biological Sciences*,

368(1614), 20120314. https://doi.org/10.1098/rstb.2012.0314

Dehning, J., Zierenberg, J., Spitzner, F. P., Wibral, M., Neto, J. P., Wilczek, M., & Priesemann, V. (2020). Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*, *369*(6500), eabb9789. https://doi.org/10.1126/science.abb9789

Dehuri, S., & Cho, S. B. (2009). Multi-criterion Pareto based particle swarm optimized polynomial neural network for classification: A review and state-of-the-art. *Computer Science Review*, *3*(1), 19–40. https://doi.org/10.1016/j.cosrev.2008.11.002

Deng, X., Tian, X., & Chen, S. (2013). Modified kernel principal component analysis based on local structure analysis and its application to nonlinear process fault diagnosis. *Chemometrics and Intelligent Laboratory Systems*, *127*, 195–209. https://doi.org/10.1016/j.chemolab.2013.07.001

Ding, J., Hostallero, D.E., El Khili, M.R., Fonseca, G.J., Milette, S., Noorah, N., Guay-Belzile, M., Spicer, J., Daneshtalab, N., Sirois, M., Tremblay, K., Emad, A., Rousseau, S. (2021). A network-informed analysis of SARS-CoV-2 and hemophagocytic lymphohistiocytosis genes' interactions points to Neutrophil extracellular traps as mediators of thrombosis in COVID-19. *PLoS Computational Biology*, *17*(3), e1008810. https://doi.org/10.1371/journal.pcbi.1008810

Dowell, A. M. (1999). Layer of protection analysis and inherently safer processes. *Process Safety Progress*, *18*(4), 214–220. https://doi.org/10.1002/prs.680180409

Du, W., Fan, Y., Zhang, Y., & Zhang, J. (2017). Fault diagnosis of non-Gaussian process based on FKICA. *Journal of the Franklin Institute*, *354*(6), 2573–2590. https://doi.org/10.1016/j.jfranklin.2016.11.012

Dureau, J., Kalogeropoulos, K., & Baguelin, M. (2013). Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, *14*(3), 541–555. https://doi.org/10.1093/biostatistics/kxs052

Eastman, B., Meaney, C., Przedborski, M., & Kohandel, M. (2021). Modeling the impact of public response on the COVID-19 pandemic in Ontario. *PLoS ONE*, *16*(4 April), e0249456. https://doi.org/10.1371/journal.pone.0249456

Engbert, R., Rabe, M. M., Kliegl, R., & Reich, S. (2021). Sequential Data Assimilation of the Stochastic SEIR Epidemic Model for Regional COVID-19 Dynamics. *Bulletin of Mathematical Biology*, *83*(1), 1–16. https://doi.org/10.1007/s11538-020-00834-8

Farah, M., Birrell, P., Conti, S., & Angelis, D. De. (2014). Bayesian Emulation and Calibration of a Dynamic Epidemic Model for A/H1N1 Influenza. *Journal of the American Statistical Association*, *109*(508), 1398–1411. https://doi.org/10.1080/01621459.2014.934453

Fast, S. M., Mekaru, S., Brownstein, J. S., Postlethwaite, T. A., & Markuzon, N. (2015). The Role of Social Mobilization in Controlling Ebola Virus in Lofa County, Liberia. *PLoS Currents*, *7*. https://doi.org/10.1371/currents.outbreaks.c3576278c66b22ab54a25e122fcdbec1

Fintzi, J., Cui, X., Wakefield, J., & Minin, V. N. (2017). Efficient Data Augmentation for Fitting Stochastic Epidemic Models to Prevalence Data. *Journal of Computational and Graphical Statistics*, *26*(4), 918–929. https://doi.org/10.1080/10618600.2017.1328365

Fuan, W., Hongkai, J., Haidong, S., Wenjing, D., & Shuaipeng, W. (2017). An adaptive deep convolutional neural network for rolling bearing fault diagnosis. *Measurement Science and Technology*, *28*(9). https://doi.org/10.1088/1361-6501/aa6e22

Galagedarage Don, M., & Khan, F. (2019). Dynamic process fault detection and diagnosis based on a combined approach of hidden Markov and Bayesian network model. *Chemical Engineering Science*, *201*, 82–96. https://doi.org/10.1016/j.ces.2019.01.060

Gan, R., Tan, J., Mo, L., Li, Y., & Huang, D. (2020). Using Partial Least Squares Regression to Fit Small Data of H7N9 Incidence Based on the Baidu Index. *IEEE Access*, *8*, 60392–60400. https://doi.org/10.1109/ACCESS.2020.2983799

Gao, Z., Cecati, C., & Ding, S. X. (2015). A survey of fault diagnosis and fault-tolerant techniques-part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, *62*(6), 3757–3767. https://doi.org/10.1109/TIE.2015.2417501

Ge, Z. (2017). Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemometrics and Intelligent Laboratory Systems*, *171*, 16–25. https://doi.org/10.1016/j.chemolab.2017.09.021

Georgieva, P., Meireles, M. J., & Feyo de Azevedo, S. (2003). Knowledge-based hybrid modelling of a batch crystallisation when accounting for nucleation, growth and agglomeration phenomena. *Chemical Engineering Science*, *58*(16), 3699–3713. https://doi.org/10.1016/S0009-2509(03)00260-4

Gharahbagheri, H., Imtiaz, S. A., & Khan, F. (2017). Root Cause Diagnosis of Process Fault Using KPCA and Bayesian Network. *Industrial and Engineering Chemistry Research*, *56*(8), 2054–2070. https://doi.org/10.1021/acs.iecr.6b01916

Ghosh, D., Moreira, J., & Mhaskar, P. (2021). Model predictive control embedding a parallel hybrid modeling strategy. *Industrial and Engineering Chemistry Research*, *60*(6), 2547–2562. https://doi.org/10.1021/acs.iecr.0c05208

Gibson, G. C., Reich, N. G., & Sheldon, D. (2020). Real-time mechanistic bayesian forecasts of COVID-19 mortality. *MedRxiv*. https://doi.org/10.1101/2020.12.22.20248736

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, *26*(6), 855–860. https://doi.org/10.1038/s41591-020-0883-7

He, S., Peng, Y., & Sun, K. (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics*, *101*(3), 1667–1680. https://doi.org/10.1007/s11071-020-05743-y

Hendriks, J., Jidling, C., Wills, A., & Schön, T. (2020). Linearly Constrained Neural Networks. arXiv:2002.01600, Retrieved from http://arxiv.org/abs/2002.01600

Hethcote, H. W. (1976). Qualitative analyses of communicable disease models. *Mathematical Biosciences*, *28*(3–4), 335–356. https://doi.org/10.1016/0025-5564(76)90132-2

Hiorns, R. W., & MacDonald, N. (1982). Time Lags in Biological Models. *Journal of the Royal Statistical Society. Series A (General)*, *145*(1), 140. https://doi.org/10.2307/2981435

Hosseini, E., Ghafoor, K. Z., Sadiq, A. S., Guizani, M., & Emrouznejad, A. (2020). COVID-19 Optimizer Algorithm, Modeling and Controlling of Coronavirus Distribution Process. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2765–2775. https://doi.org/10.1109/JBHI.2020.3012487

House, T., & Keeling, M. J. (2008). Deterministic epidemic models with explicit household structure. *Mathematical Biosciences*, *213*(1), 29–39. https://doi.org/10.1016/j.mbs.2008.01.011

Hu, Z., Cui, Q., Han, J., Wang, X., Sha, W. E. I., & Teng, Z. (2020). Evaluation and prediction of the COVID-19 variations at different input population and quarantine strategies, a case study in Guangdong province, China. *International Journal of Infectious Diseases*, *95*, 231–240. https://doi.org/10.1016/j.ijid.2020.04.010

Imtiaz, S. A., & Shah, S. L. (2008). Treatment of missing values in process data analysis. *Canadian Journal of Chemical Engineering*, *86*(5), 838–858. https://doi.org/10.1002/cjce.20099

Isa Irawan, M., & Amiroch, S. (2015). Construction of phylogenetic tree using neighbor joining

algorithms to identify the host and the spreading of SARS epidemic. *Journal of Theoretical and Applied Information Technology*, *71*(3), 424–429.

Isermann, R. (2005). Model-based fault-detection and diagnosis - Status and applications. *Annual Reviews in Control*, *29*(1), 71–85. https://doi.org/10.1016/j.arcontrol.2004.12.002

Ivorra, B., Ferrández, M. R., Vela-Pérez, M., & Ramos, A. M. (2020). Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Communications in Nonlinear Science and Numerical Simulation*, *88*, 105303. https://doi.org/10.1016/j.cnsns.2020.105303

Jiao, Z., Hu, P., Xu, H., & Wang, Q. (2020). Machine learning and deep learning in chemical health and safety: A systematic review of techniques and applications. *Journal of Chemical Health and Safety*. https://doi.org/10.1021/acs.chas.0c00075

Jin, X., Wang, S., Huang, B., & Forbes, F. (2012). Multiple model based LPV soft sensor development with irregular/missing process output measurement. *Control Engineering Practice*, *20*(2), 165–172. https://doi.org/10.1016/j.conengprac.2011.10.007

Juang, C. F. (2004). A Hybrid of Genetic Algorithm and Particle Swarm Optimization for Recurrent Network Design. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(2), 997–1006. https://doi.org/10.1109/TSMCB.2003.818557

Jung, E., Lee, J., & Chowell, G. (2016). Evaluating the number of sickbeds during ebola epidemics using optimal control theory. In *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases* (pp. 89–101). https://doi.org/10.1007/978-3-319-40413-4_7

Kao, Y. H., & Eisenberg, M. C. (2018). Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment. *Epidemics*, *25*, 89–100. https://doi.org/10.1016/j.epidem.2018.05.010

Kaspar, M. H., & Harmon Ray, W. (1993). Dynamic PLS modelling for process control. *Chemical Engineering Science*, *48*(20), 3447–3461. https://doi.org/10.1016/0009-2509(93)85001-6

Kermack, W.O., McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *115*(772), 700–721. https://doi.org/10.1098/rspa.1927.0118

Khan, A., Hussain, G., Zahri, M., Zaman, G., & Wannasingha Humphries, U. (2020). A stochastic SACR epidemic model for HBV transmission. *Journal of Biological Dynamics*, *14*(1), 788–801. https://doi.org/10.1080/17513758.2020.1833993

Khatibisepehr, S., & Huang, B. (2008). Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Industrial and Engineering Chemistry Research*, *47*(22), 8713–8723. https://doi.org/10.1021/ie800386v

Kim, S., Ko, Y., Kim, Y. J., & Jung, E. (2020). The impact of social distancing and public behavior changes on COVID-19 transmission dynamics in the Republic of Korea. *PLoS ONE*, *15*(9 September 2020), e0238684. https://doi.org/10.1371/journal.pone.0238684

Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, *28*(1), 3–21. https://doi.org/10.1016/0169-7439(95)80036-9

Kozioł, K., Stanisławski, R., & Bialic, G. (2020). Fractional-order sir epidemic model for transmission prediction of COVID-19 disease. *Applied Sciences (Switzerland)*, *10*(23), 1–9. https://doi.org/10.3390/app10238316

Kumar, A., Goel, K., & Nilam. (2020). A deterministic time-delayed SIR epidemic model: mathematical modeling and analysis. *Theory in Biosciences*, *139*(1), 67–76. https://doi.org/10.1007/s12064-019-00300-7

Kypraios, T., Neal, P., & Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. *Mathematical Biosciences*, *287*, 42–53. https://doi.org/10.1016/j.mbs.2016.07.001

Lange, A. (2016). Reconstruction of disease transmission rates: Applications to measles, dengue, and influenza. *Journal of Theoretical Biology*, *400*, 138–153. https://doi.org/10.1016/j.jtbi.2016.04.017

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, *6*, 1783–1816.

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote Sensing*, *67*(10), 1137–1142.

Lawson, A. B. (2020). NIMBLE for Bayesian Disease Mapping. *Spatial and Spatio-Temporal Epidemiology*, *33*, 100323. https://doi.org/10.1016/j.sste.2020.100323

Lee, J., & Kang, S. (2007). GA based meta-modeling of BPN architecture for constrained approximate optimization. *International Journal of Solids and Structures*, *44*(18), 5980–5993. https://doi.org/10.1016/j.ijsolstr.2007.02.008

Lee, S. Y., Lei, B., & Mallick, B. (2020). Estimation of COVID-19 spread curves integrating global data and borrowing information. *PLoS ONE*, *15*(7 July), 1–17. https://doi.org/10.1371/journal.pone.0236860

Lee, V. J., Lye, D. C., & Wilder-Smith, A. (2009). Combination strategies for pandemic influenza response - a systematic review of mathematical modeling studies. *BMC Medicine*, *7,* 76. https://doi.org/10.1186/1741-7015-7-76

Legrand, J., Grais, R. F., Boelle, P. Y., Valleron, A. J., & Flahault, A. (2007). Understanding the dynamics of Ebola epidemics. *Epidemiology and Infection*, *135*(4), 610–621. https://doi.org/10.1017/S0950268806007217

Lekone, P. E., & Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, *62*(4), 1170–1177. https://doi.org/10.1111/j.1541-0420.2006.00609.x

Li, G., Qin, S. J., & Yuan, T. (2016). Data-driven root cause diagnosis of faults in process industries. *Chemometrics and Intelligent Laboratory Systems*, *159*, 1–11. https://doi.org/10.1016/j.chemolab.2016.09.006

Li, M., Dushoff, J., & Bolker, B. M. (2018). Fitting mechanistic epidemic models to data: A comparison of simple Markov chain Monte Carlo approaches. *Statistical Methods in Medical Research*, *27*(7), 1956–1967. https://doi.org/10.1177/0962280217747054

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, *231*, 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114

Lin, Q., Zhao, S., Gao, D., Lou, Y., Yang, S., Musa, S.S., Wang, M.H., Cai, Y., Wang, W., Yang, L., He, D. (2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *International Journal of Infectious Diseases*, *93*, 211–216. https://doi.org/10.1016/j.ijid.2020.02.058

Lindhout, P., & Reniers, G. (2020). Reflecting on the safety zoo: Developing an integrated pandemics barrier model using early lessons from the Covid-19 pandemic. *Safety science*, *130*, 104907. https://doi.org/10.1016/j.ssci.2020.104907

Lipsitch, M., Cohen, T., Cooper, B., Robins, J. M., Ma, S., James, L., Gopalakrishna, G., Chew, S. K., Tan, C. C., Samore, M.H., Fisman, D., Murray, M. (2003). Transmission dynamics and

control of severe acute respiratory syndrome. *Science*, *300*(5627), 1966–1970. https://doi.org/10.1126/science.1086616

Liu, S., McDermid, J. A., & Chen, Y. (2010). A rigorous method for inspection of model-Based formal specifications. *IEEE Transactions on Reliability*, *59*(4), 667–684. https://doi.org/10.1109/TR.2010.2085571

Lytras, T., Gkolfinopoulou, K., Bonovas, S., & Nunes, B. (2019). FluHMM: A simple and flexible Bayesian algorithm for sentinel influenza surveillance and outbreak detection. *Statistical Methods in Medical Research*, *28*(6), 1826–1840. https://doi.org/10.1177/0962280218776685

Mallick, M. R., & Imtiaz, S. A. (2013). A hybrid method for process fault detection and diagnosis. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, *10*(PART 1), 827–832. https://doi.org/10.3182/20131218-3-IN-2045.00099

Manevski, D., Ružić Gorenjec, N., Kejžar, N., & Blagus, R. (2020). Modeling COVID-19 pandemic using Bayesian analysis with application to Slovene data. *Mathematical Biosciences*, 329, 108466. https://doi.org/10.1016/j.mbs.2020.108466

Maronna, R. A. (1976). Robust M-Estimators of Multivariate Location and Scatter. *The Annals of Statistics*, *4*(1), 51–67. https://doi.org/10.1214/aos/1176343347

Martínez-Beneito, M. A., Conesa, D., López-Quílez, A., & López-Maside, A. (2008). Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine*, *27*(22), 4455–4468. https://doi.org/10.1002/sim.3320

Marwa, Y. M., Mwalili, S., & Mbalawata, I. S. (2019). Bayesian approach in modelling cholera outbreak in Ilala municipal council, Tanzania. *Journal of Mathematical and Computational Science*, *9*(1), 73–86. https://doi.org/10.28919/jmcs/3881

McKinley, T.J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J.E., Nsubuga, R.N., Goldstein, M., White, R.G. (2018). Approximate bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statistical Science*, *33*(1), 4–18. https://doi.org/10.1214/17-STS618

Minter, A., & Retkute, R. (2019). Approximate Bayesian Computation for infectious disease modelling. *Epidemics*, *29*(August), 100368. https://doi.org/10.1016/j.epidem.2019.100368

Monteiro, L. H. A., Gandini, D. M., & Schimit, P. H. T. (2020). The influence of immune individuals in disease spread evaluated by cellular automaton and genetic algorithm.

Computer Methods and Programs in Biomedicine, *196*, 105707. https://doi.org/10.1016/j.cmpb.2020.105707

Muralidhar, N., Islam, M. R., & Marwah, M. (2018). Incorporating Prior Domain Knowledge into Deep Neural Networks. *IEEE Big Data*, 36–45. Retrieved from http://people.cs.vt.edu/naren/papers/PID5657885.pdf

Nawi, N. M., Khan, A., & Rehman, M. Z. (2013). A New Levenberg Marquardt based Back Propagation Algorithm Trained with Cuckoo Search. *Procedia Technology*, *11*(Iceei), 18–23. https://doi.org/10.1016/j.protcy.2013.12.157

Neal, P. J. (2019). Approximate Bayesian Computation Methods for Epidemic Models. In *Handbook of Infectious Disease Data Analysis* (pp. 179–197). https://doi.org/10.1201/9781315222912-10

Neal, P., & Terry Huang, C. L. (2015). Forward Simulation Markov Chain Monte Carlo with Applications to Stochastic Epidemic Models. *Scandinavian Journal of Statistics*, *42*(2), 378–396. https://doi.org/10.1111/sjos.12111

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences*, *180*(1–2), 103–114. https://doi.org/10.1016/S0025-5564(02)00109-8

Okyere, E., De-Graft Ankamah, J., Hunkpe, A. K., & Mensah, D. (2020). Deterministic Epidemic Models for Ebola Infection with Time-Dependent Controls. *Discrete Dynamics in Nature and Society*, *2020*, 1–12. https://doi.org/10.1155/2020/2823816

Oliveira, R. (2004). Combining first principles modelling and artificial neural networks: A general framework. In *Computers and Chemical Engineering* (Vol. 28, pp. 755–766). https://doi.org/10.1016/j.compchemeng.2004.02.014

Osei, F. B., Duker, A. A., & Stein, A. (2012). Bayesian structured additive regression modeling of epidemic data: Application to cholera. *BMC Medical Research Methodology*, *12*. https://doi.org/10.1186/1471-2288-12-118

Otoo, D., Opoku, P., Charles, S., & Kingsley, A. P. (2020). Deterministic epidemic model for (SVCSyCAsyIR) pneumonia dynamics, with vaccination and temporal immunity. *Infectious Disease Modelling*, *5*, 42–60. https://doi.org/10.1016/j.idm.2019.11.001

Paiva, H. M., Afonso, R. J. M., de Oliveira, I. L., & Garcia, G. F. (2020). A data-driven model to describe and forecast the dynamics of COVID-19 transmission. *PLoS ONE*, *15*(7), e0236386.

https://doi.org/10.1371/journal.pone.0236386

Pedroso-Rodriguez, L. M., Marrero, A., & De Arazoza, H. (2003). Nonlinear parametric model identification using genetic algorithms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2687*, 473–480. https://doi.org/10.1007/3-540-44869-1_60

Plummer, M. (2003). JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling JAGS : Just Another Gibbs Sampler. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing; Vienna, Austria*, (Dsc).

Pollicott, M., Wang, H., & Weiss, H. (2012). Extracting the time-dependent transmission rate from infection data via solution of an inverse ODE problem. *Journal of Biological Dynamics*, *6*(2), 509–523. https://doi.org/10.1080/17513758.2011.645510

Polo, G., Mera Acosta, C., Soler-Tovar, D., Porras Villamil, J.F., Palencia, N.P., Penagos, M., Meza Martinez, J., Bobadilla, J.N., Martin, L.V., Durán, S., Rodriguez Álvarez, M., Meza Carvajalino, C., Villamil, L.C., Benavides Ortiz, E. (2020). Bayesian Spatio-Temporal Modeling of COVID-19: Inequalities on Case-Fatality Risk. *MedRxiv*. https://doi.org/10.1101/2020.08.18.20171074

Ponciano, J. M., & Capistrán, M. A. (2011). First principles modeling of nonlinear incidence rates in seasonal epidemics. *PLoS Computational Biology*, *7*(2), e1001079. https://doi.org/10.1371/journal.pcbi.1001079

Popinga, A., Vaughan, T., Stadler, T., & Drummond, A. J. (2014). Inferring epidemiological dynamics with Bayesian coalescent inference: The merits of deterministic and stochastic models. *Genetics*, *199*(2), 595–607. https://doi.org/10.1534/genetics.114.172791

Price, D. J., Bean, N. G., Ross, J. V., & Tuke, J. (2016). On the efficient determination of optimal Bayesian experimental designs using ABC: A case study in optimal observation of epidemics. *Journal of Statistical Planning and Inference*, *172*, 1–15. https://doi.org/10.1016/j.jspi.2015.12.008

Psichogios, D. C., & Ungar, L. H. (1992). A hybrid neural network-first principles approach to process modeling. *AIChE Journal, 38(10), 1499-1511*. https://doi.org/10.1002/aic.690381003

Qin, S. J. (1998). Recursive PLS algorithms for adaptive data modeling. *Computers and Chemical Engineering*, *22*(4), 503–514. https://doi.org/10.1016/s0098-1354(97)00262-7

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control, 36(2),* 220–234. https://doi.org/10.1016/j.arcontrol.2012.09.004

Rachah, A., & Torres, D. F. M. (2016). Dynamics and Optimal Control of Ebola Transmission. *Mathematics in Computer Science*, *10*(3), 331–342. https://doi.org/10.1007/s11786-016-0268-y

Rashid, M. M., & Yu, J. (2012). Hidden markov model based adaptive independent component analysis approach for complex chemical process monitoring and fault detection. *Industrial and Engineering Chemistry Research*, *51*(15), 5506–5514. https://doi.org/10.1021/ie300203u

Ren M, Zeng W, Yang B, Urtasun R. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning 2018* Jul 3 (pp. 4334-4343). *PMLR.*

Ren, X., Zhu, K., Cai, T., & Li, S. (2017). Fault Detection and Diagnosis for Nonlinear and Non-Gaussian Processes Based on Copula Subspace Division. *Industrial and Engineering Chemistry Research*, *56*(40), 11545–11564. https://doi.org/10.1021/acs.iecr.7b02419

Reshef, Y. A., Reshef, D. N., Sabeti, P. C., & Mitzenmacher, M. (2014). Theoretical Foundations of Equitability and the Maximal Information Coefficient. *ArXiv Preprint ArXiv:1408.4908*. Retrieved from http://arxiv.org/abs/1408.4908

Rivers, C. M., Lofgren, E. T., Marathe, M., Eubank, S., & Lewis, B. L. (2014). Modeling the Impact of Interventions on an Epidemic of Ebola in Sierra Leone and Liberia. *PLoS Currents, 6*. https://doi.org/10.1371/currents.outbreaks.4d41fe5d6c05e9df30ddce33c66d084c

Rizzo, C., & Atti, M. L. C. degli. (2008). Modeling influenza pandemic and interventions. In *Influenza Vaccines for the Future* (pp. 281–296). https://doi.org/10.1007/978-3-7643-8371-8_13

Romero-Severson, E. O., Ribeiro, R. M., & Castro, M. (2018). Noise is not error: Detecting parametric heterogeneity between epidemiologic time series. *Frontiers in Microbiology*, *9*, 1529. https://doi.org/10.3389/fmicb.2018.01529

Russell, T.W., Golding, N., Hellewell, J., Abbott, S., Wright, L., Pearson, C.A.B., van Zandvoort, K., Jarvis, C.I., Gibbs, H., Liu, Y., Eggo, R.M., Edmunds, W.J., Kucharski, A.J., Deol, A.K., Villabona-Arenas, C.J., Jombart, T., O'Reilly, K., Munday, J.D., Meakin, S.R., Lowe, R., Gimma, A., Endo, A., Nightingale, E.S., Medley, G., Foss, A.M., Knight, G.M., Prem, K., Hué, S., Diamond, C., Rudge, J.W., Atkins, K.E., Auzenbergs, M., Flasche, S., Houben, R.M.G.J., Quilty, B.J., Klepac, P., Quaife, M., Funk, S., Leclerc, Q.J., Emery, J.C., Jit, M.,

Simons, D., Bosse, N.I., Procter, S.R., Sun, F.Y., Clifford, S., Sherratt, K., Rosello, A., Davies, N.G., Brady, O., Tully, D.C., Gore-Langton, G.R. (2020). Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. *BMC Medicine*, *18*(1), 332. https://doi.org/10.1186/s12916-020-01790-9

Safari, S., Shabani, F., & Simon, D. (2014). Multirate multisensor data fusion for linear systems using Kalman filters and a neural network. *Aerospace Science and Technology*, *39*, 465–471. https://doi.org/10.1016/j.ast.2014.06.005

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Seheult, A. H., Green, P. J., Rousseeuw, P. J., & Leroy, A. M. (1989). Robust Regression and Outlier Detection. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *152*(1), 133. https://doi.org/10.2307/2982847

Shangguan, D., Liu, Z., Wang, L., & Tan, R. (2021). A stochastic epidemic model with infectivity in incubation period and homestead–isolation on the susceptible. *Journal of Applied Mathematics and Computing*, 1–21. https://doi.org/10.1007/s12190-021-01504-1

Sharkey, K. J. (2011). Deterministic epidemic models on contact networks: Correlations and unbiological terms. *Theoretical Population Biology*, *79*(4), 115–129. https://doi.org/10.1016/j.tpb.2011.01.004

Sharmin, S., Glass, K., Viennet, E., & Harley, D. (2018). A Bayesian approach for estimating under-reported dengue incidence with a focus on non-linear associations between climate and dengue in Dhaka, Bangladesh. *Statistical Methods in Medical Research*, *27*(4), 991–1000. https://doi.org/10.1177/0962280216649216

Shen, M., Xiao, Y., & Rong, L. (2015). Modeling the effect of comprehensive interventions on Ebola virus transmission. *Scientific Reports*, *5,* 15818. https://doi.org/10.1038/srep15818

Smirnova, A., deCamp, L., & Chowell, G. (2019). Forecasting Epidemics Through Nonparametric Estimation of Time-Dependent Transmission Rates Using the SEIR Model. *Bulletin of Mathematical Biology*, *81*(11), 4343–4365. https://doi.org/10.1007/s11538-017-0284-3

Smyth, P. (1994). Hidden Markov models for fault detection in dynamic systems. *Pattern Recognition*, *27*(1), 149–164. https://doi.org/10.1016/0031-3203(94)90024-8

Staroswiecki, M. (2000). Quantitative and qualitative models for fault detection and isolation. *Mechanical Systems and Signal Processing*, *14*(3), 301–325. https://doi.org/10.1006/mssp.2000.1293

Stefatos, G., & Hamza, A. Ben. (2010). Dynamic independent component analysis approach for fault detection and diagnosis. *Expert Systems with Applications*, *37*(12), 8606–8617. https://doi.org/10.1016/j.eswa.2010.06.101

Stewart, R., & Ermon, S. (2017). Label-free supervision of neural networks with physics and domain knowledge. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, *1*(1), 2576–2582.

Subramanian, R., He, Q., & Pascual, M. (2021). Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9), e2019716118. https://doi.org/10.1073/pnas.2019716118

Subudhi, B., & Jena, D. (2011a). A differential evolution based neural network approach to nonlinear system identification. *Applied Soft Computing Journal*, *11*(1), 861–871. https://doi.org/10.1016/j.asoc.2010.01.006

Subudhi, B., & Jena, D. (2011b). Nonlinear system identification using memetic differential evolution trained neural networks. *Neurocomputing*, *74*(10), 1696–1709. https://doi.org/10.1016/j.neucom.2011.02.006

Taghizadeh, L., Karimi, A., & Heitzinger, C. (2020). Uncertainty quantification in epidemiological models for the COVID-19 pandemic. *Computers in Biology and Medicine*, *125,* 104011. https://doi.org/10.1016/j.compbiomed.2020.104011

Tan, K. C., & Li, Y. (2002). Grey-box model identification via evolutionary computing. *Control Engineering Practice*, *10*(7), 673–684. https://doi.org/10.1016/S0967-0661(02)00031-X

Taylor, B. P., Dushoff, J., & Weitz, J. S. (2016). Stochasticity and the limits to confidence when estimating R0 of Ebola and other emerging infectious diseases. *Journal of Theoretical Biology*, *408*, 145–154. https://doi.org/10.1016/j.jtbi.2016.08.016

Thissen, U., Swierenga, H., De Weijer, A. P., Wehrens, R., Melssen, W. J., & Buydens, L. M. C. (2005). Multivariate statistical process control using mixture modelling. *Journal of Chemometrics*, *19*(1), 23–31. https://doi.org/10.1002/cem.903

Thompson, M. L., & Kramer, M. A. (1994). Modeling chemical processes using prior knowledge

and neural networks. *AIChE Journal, 40(8), 1328-1340*. https://doi.org/10.1002/aic.690400806

Tidriri, K., Chatti, N., Verron, S., & Tiplica, T. (2016). Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control, 42, 63-81*. https://doi.org/10.1016/j.arcontrol.2016.09.008

Tong, C., Palazoglu, A., & Yan, X. (2014). Improved ICA for process monitoring based on ensemble learning and Bayesian inference. *Chemometrics and Intelligent Laboratory Systems*, *135*, 141–149. https://doi.org/10.1016/j.chemolab.2014.04.012

Touloupou, P., Finkenstädt, B., & Spencer, S. E. F. (2020). Scalable Bayesian Inference for Coupled Hidden Markov and Semi-Markov Models. *Journal of Computational and Graphical Statistics*, *29*(2), 238–249. https://doi.org/10.1080/10618600.2019.1654880

Trawicki, M. B. (2017). Deterministic seirs epidemic model for modeling vital dynamics, vaccinations, and temporary immunity. *Mathematics*, *5*(1), 1–7. https://doi.org/10.3390/math5010007

Tritch, W., & Allen, L. J. S. (2018). Duration of a minor epidemic. *Infectious Disease Modelling*, *3*, 60–73. https://doi.org/10.1016/j.idm.2018.03.002

Valvo, P. S. (2020). A bimodal lognormal distribution model for the prediction of COVID-19 deaths. *Applied Sciences (Switzerland)*, *10*(23), 1–24. https://doi.org/10.3390/app10238500

Van Lith, P. F., Betlem, B. H. L., & Roffel, B. (2003). Combining prior knowledge with data driven modeling of a batch distillation column including start-up. *Computers and Chemical Engineering*, *27*(7), 1021–1030. https://doi.org/10.1016/S0098-1354(03)00067-X

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis part I: Quantitative model-based methods. *Computers and Chemical Engineering*, *27*(3), 293–311. https://doi.org/10.1016/S0098-1354(02)00160-6

Verron, S., Tiplica, T., & Kobi, A. (2007). Fault diagnosis of industrial systems with bayesian networks and mutual information. In *2007 European Control Conference, ECC 2007* (pp. 2304–2311). Nancy. https://doi.org/10.23919/ecc.2007.7068252

Wang, W., Ji, C., Bi, Y., & Liu, S. (2020). Stability and asymptoticity of stochastic epidemic model with interim immune class and independent perturbations. *Applied Mathematics Letters*, *104*, 106245. https://doi.org/10.1016/j.aml.2020.106245

White, L. F., & Pagano, M. (2008). A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine*, *27*(16), 2999–3016. https://doi.org/10.1002/sim.3136

Widodo, A., & Yang, B. S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *21*(6), 2560–2574. https://doi.org/10.1016/j.ymssp.2006.12.007

Wu, H., & Zhao, J. (2018). Deep convolutional neural network model based chemical process fault diagnosis. *Computers and Chemical Engineering*, *115*, 185–197. https://doi.org/10.1016/j.compchemeng.2018.04.009

Wu, K. M., & Riley, S. (2016). Estimation of the basic reproductive number and mean serial interval of a novel pathogen in a small, well-observed discrete population. *PLoS ONE*, *11*(2), e0148061. https://doi.org/10.1371/journal.pone.0148061

Wu, Z., Rincon, D., & Christofides, P. D. (2020). Process structure-based recurrent neural network modeling for model predictive control of nonlinear processes. *Journal of Process Control*, *89*, 74–84. https://doi.org/10.1016/j.jprocont.2020.03.013

Xiong, Q., & Jutan, A. (2002). Grey-box modelling and control of chemical processes. *Chemical Engineering Science*, *57*(6), 1027–1039. https://doi.org/10.1016/S0009-2509(01)00439-0

Xu, X., Kypraios, T., & O'Neill, P. D. (2016). Bayesian non-parametric inference for stochastic epidemic models using Gaussian Processes. *Biostatistics*, *17*(4), 619–633. https://doi.org/10.1093/biostatistics/kxw011

Yan, X., & Zou, Y. (2008). Optimal and sub-optimal quarantine and isolation control in SARS epidemics. *Mathematical and Computer Modelling*, *47*(1), 235–245. https://doi.org/10.1016/j.mcm.2007.04.003

Yang, F., Dai, C., Tang, J., Xuan, J., & Cao, J. (2020). A hybrid deep learning and mechanistic kinetics model for the prediction of fluid catalytic cracking performance. *Chemical Engineering Research and Design*, *155*, 202–210. https://doi.org/10.1016/j.cherd.2020.01.013

Yin, S., Ding, S. X., Haghani, A., Hao, H., & Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control*, *22*(9), 1567–1581. https://doi.org/10.1016/j.jprocont.2012.06.009

Yin, S., Ding, S. X., Xie, X., & Luo, H. (2014). A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics*, 61(11), 6418-6428. https://doi.org/10.1109/TIE.2014.2301773

Yu, H., Khan, F., & Garaniya, V. (2015a). A probabilistic multivariate method for fault diagnosis of industrial processes. *Chemical Engineering Research and Design*, *104*(3), 306–318. https://doi.org/10.1016/j.cherd.2015.08.026

Yu, H., Khan, F., & Garaniya, V. (2015b). Nonlinear Gaussian Belief Network based fault diagnosis for industrial processes. *Journal of Process Control*, *35*, 178–200. https://doi.org/10.1016/j.jprocont.2015.09.004

Yu, H., Khan, F., & Garaniya, V. (2016). An Alternative Formulation of PCA for Process Monitoring Using Distance Correlation. *Industrial and Engineering Chemistry Research*, *55*(3), 656–669. https://doi.org/10.1021/acs.iecr.5b03397

Yu, J. (2012a). A nonlinear kernel Gaussian mixture model based inferential monitoring approach for fault detection and diagnosis of chemical processes. *Chemical Engineering Science*, *68*(1), 506–519. https://doi.org/10.1016/j.ces.2011.10.011

Yu, J. (2012b). Multiway discrete hidden Markov model-based approach for dynamic batch process monitoring and fault classification. *AIChE Journal*, *58*(9), 2714–2725. https://doi.org/10.1002/aic.12794

Yu, J. (2013). A new fault diagnosis method of multimode processes using Bayesian inference based Gaussian mixture contribution decomposition. *Engineering Applications of Artificial Intelligence*, *26*(1), 456–466. https://doi.org/10.1016/j.engappai.2012.09.003

Yu, J., & Qin, S. J. (2008). Multimode process monitoring with bayesian inference-based finite Gaussian mixture models. *AIChE Journal*, *54*(7), 1811–1829. https://doi.org/10.1002/aic.11515

Yu, J., & Qin, S. J. (2009). Multiway gaussian mixture model based multiphase batch process monitoring. *Industrial and Engineering Chemistry Research*, *48*(18), 8585–8594. https://doi.org/10.1021/ie900479g

Yu, J., & Rashid, M. M. (2013). A novel dynamic bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis. *AIChE Journal*, *59*(7), 2348–2365. https://doi.org/10.1002/aic.14013

Yuan, X., Wan, Y., Yan, C., Ge, Z., Song, Z., & Gui, W. (2017). Weighted linear dynamic system

for feature representation and soft sensor application in nonlinear dynamic industrial processes. *IEEE Transactions on Industrial Electronics*, *65*(2), 1508–1517. https://doi.org/10.1109/TIE.2017.2733443

Zhang, S., Bi, K., & Qiu, T. (2020). Bidirectional Recurrent Neural Network-Based Chemical Process Fault Diagnosis. *Industrial and Engineering Chemistry Research*, *59*(2), 824–834. https://doi.org/10.1021/acs.iecr.9b05885

Zhang, Y., & Qin, S. J. (2007). Fault detection of nonlinear processes using multiway kernel independent component analysis. *Industrial and Engineering Chemistry Research*, *46*(23), 7780–7787. https://doi.org/10.1021/ie070381q

Zhang, Y., & Zhang, Y. (2010). Fault detection of non-Gaussian processes based on modified independent component analysis. *Chemical Engineering Science*, *65*(16), 4630–4639. https://doi.org/10.1016/j.ces.2010.05.010

Zhang, Z., & Dong, F. (2014). Fault detection and diagnosis for missing data systems with a three time-slice dynamic Bayesian network approach. *Chemometrics and Intelligent Laboratory Systems*, *138*, 30–40. https://doi.org/10.1016/j.chemolab.2014.07.009

Zhao, H., Sun, S., & Jin, B. (2018). Sequential Fault Diagnosis Based on LSTM Neural Network. *IEEE Access*, *6*, 12929–12939. https://doi.org/10.1109/ACCESS.2018.2794765

Zhao, Hui, Liu, J., Dong, W., Sun, X., & Ji, Y. (2017). An improved case-based reasoning method and its application on fault diagnosis of Tennessee Eastman process. *Neurocomputing*, *249*, 266–276. https://doi.org/10.1016/j.neucom.2017.04.022

Zhou, D., Li, G., & Qin, S. J. (2010). Total projection to latent structures for process monitoring. *AIChE Journal*, *56*(1), 168–178. https://doi.org/10.1002/aic.11977

Zhou, X., Li, X., & Wang, W. S. (2014). Bifurcations for a deterministic SIR epidemic model in discrete time. *Advances in Difference Equations*, *2014*(1), 168. https://doi.org/10.1186/1687-1847-2014-168

Zhu, J., Shi, H., Song, B., Tao, Y., & Tan, S. (2020). Information concentrated variational auto-encoder for quality-related nonlinear process monitoring. *Journal of Process Control*, *94*, 12–25. https://doi.org/10.1016/j.jprocont.2020.08.002

Chapter 3

Quality based training (QbT) of neural networks for robust fault detection in the presence of mislabeled data

Preface*: The success of the data-based models is dependent on the availability of reliable data for developing the models. Process systems frequently encounter low quality, inconsistent, sparse, noisy, and imbalanced data. Thus, robust models need to be devised for handling these inconsistencies. This presents a robust data-driven model for the safety of process systems by harnessing data-quality features in the training of the diagnostic networks. This work can be mapped to the sub-objective "development of robust data-driven models for the safety of process systems" of this thesis.*

Bibliographic citation and authorship statement: The bibliographic citation and contributions of authors and co-authors are presented below. The first author (Md Alauddin) was responsible for simulation, analysis and writing of the manuscript. The co-authors helped in developing the manuscript through conceptualization, development, writing, review & editing, supervision, project administration, and funding acquisition.

Alauddin, M., Khan, F., Imtiaz, S., Ahmad, S., & Amyotte, P., Quality-based training (QbT) of supervised neural networks for robust fault detection in the presence of mislabeled data, *(draft under the committee review)*

> *Md Alauddin**: Formal Analysis, Methodology, Software; Investigation, Validation; Writing - Original Draft; Writing - Review & Editing*
>
> *Faisal Khan: Conceptualization, Methodology, Writing - Review & Editing; Supervision; Project administration; Funding acquisition*
>
> *Syed Imtiaz: Conceptualization, Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition*
>
> *Salim Ahmed: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition*
>
> *Paul Amyotte: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition*

Abstract: This work presents a novel approach, the quality-based training (QbT) of the supervised neural networks, for handling mislabeled and/or low-quality data. The approach is based on explicitly addressing the quality issues of a given dataset and assigning a lower weight to the low-quality data. The method has been examined on two case studies; a continuous stirred tank heater problem and the Tennessee Eastman chemical process. The performance of the proposed robust ANN is evaluated in terms of the fault detection rate of the legitimate and incorrectly labeled data. A more robust classification index has also been developed to evaluate the fault detection systems. The proposed robust ANN yields improved results on legitimate data. The proposed model also exhibits robustness by maintaining higher performance at 1%, 5%, and 10% mislabeled training and testing data where the performance of the conventional ANN is severely affected. The proposed methodology is advantageous in handling low-quality data in detecting faults of process systems.

Keywords: fault detection, fault diagnosis, process monitoring, fault detection rate, false alarm rate.

3.1 Introduction:

Data are being referred to as today's *digital oil* (Yi, Liu, Liu, & Jin, 2014), the *new raw material of the 21st century* (Berners-Lee & Shadbolt, 2011), and *the world's most valuable resource* (Jossen, 2017). Process data contain valuable information that can be mined to characterize uncertain and emerging situations that are not considered in the process design phase. The advancement in big data has prompted many industries, including process industries, to re-examine their traditional roles for design, control, and maintenance using data-based systems. It has led to a growing consensus towards intelligent and data-driven operations for predictive maintenance in processing systems using process monitoring, fault detection, fault diagnosis, fault characterization, accident modeling, and risk management. Numerous data-based methodologies such as principal component analysis (PCA) and partial least squares (PLS) (Kaspar & Harmon Ray, 1993; Kourti

& MacGregor, 1995; Qin, 1998), independent component analysis (ICA) (Stefatos & Hamza, 2010; Zhang & Qin, 2007; Zhang & Zhang, 2010), vine copula-based methods (Cui & Li, 2020; Ren et al., 2017; Yu, Khan & Garaniya, 2015a), Gaussian mixture model (GMM) (Yu, 2013; Yu & Qin, 2009; Li, Qin & Yuan, 2016; Thissen et al., 2005; Yu & Qin, 2008) have been studied for fault detection and diagnosis. Many machine learning techniques such as artificial neural network (ANN) (Agatonovic-Kustrin & Beresford, 2000; Li et al., 2017; Park & Sandberg, 1991; Schmidhuber, 2015; Yu, Qu, Gao, & Tian, 2019; Zhang, Bi, & Qiu, 2020; Zhao, Sun, & Jin, 2018), support vector machines (SVM) (Burges, 1998; Cortes & Vapnik, 1995; Widodo & Yang, 2007), case-based reasoning (CBR) (Zhao et al., 2017), and classification and regression tree (CART) (Lawrence & Wright, 2001) have been successfully used for development of fault detection and diagnosis (FDD) methods. Alauddin, Khan, Imtiaz, and Ahmed (2018) classified the evolution and development of research on data-driven FDD in four categories, namely, formulation of basic data-driven algorithms, advancement of the algorithms, applications of the data-based algorithms in process systems, and development and application of advanced hybrid techniques in the process systems.

Several comprehensive reviews on data-driven FDD have been published (Alauddin et al., 2018; Arunthavanathan, Khan, Ahmed, & Imtiaz, 2021; Dai & Gao, 2013; Gao, Cecati, & Ding, 2015; Ge, 2017; Qin, 2012; Tidriri, Chatti, Verron, & Tiplica, 2016; Yin, Ding, Haghani, Hao, & Zhang, 2012; Yin, Ding, Xie, & Luo, 2014).

Tremendous process measurement data are available in industrial systems because of today's smart sensors. However, selecting legitimate data for supervisory training in process systems is still a daunting task. Frequent data-quality problems are incomplete data, outliers, imbalanced data, incorrectly mapped properties, noise, and inconsistency. The common causes of low-quality data

in a process system include sensor breakdown, process shutdown, malfunctioning of equipment, random fluctuations, incorrect calibration, inconsistent sampling frequencies, and data entry errors due to human factors (Imtiaz & Shah, 2008). The imbalance and sparse data are also a growing concern for efficient training of data-driven models of process systems having a high volume of data for the normal conditions while limited data representing the anomalies. The low-quality data affect model parameters, resulting in inferior process monitoring performance leading to a failure of the intended function. Accordingly, robust models are required for dealing with low-quality data of process systems.

Many robust techniques have been devised to deal with low-quality data. The missing data problem is handled by mean substitution, single regression imputation, last observation carried forward (LOCF), as well as methods based on maximum likelihood and Bayesian multiple imputations (Imtiaz & Shah, 2008; Jin, Wang, Huang, & Forbes, 2012; Khatibisepehr & Huang, 2008; Schafer & Graham, 2002). Bayesian network (BN) and hidden Markov model (HMM) are efficient tools based on probabilistic reasoning for handling uncertainty. The HMM estimates the probability distributions of state transitions and unobservable states of the process. Independent component analysis (ICA) was developed to handle the non-Gaussian data that are frequently encountered in process systems. Rashid and Yu (2012) introduced the concept of hidden Markov's model in ICA for handling multimodal problems of industrial processes. Cai and Tian (2014) and Cai, Tian, and Chen (2014) proposed robust ICA for handling outliers and noisy data. Vine copula-based methods are other potential tools for handling non-Gaussian data. Unlike PCA and PLS which are based on dimensionality reduction, the vine copula-based FDD tools exploit features correlation for calculating the copula density and joint probability distribution for estimating the fault probability.

Robust statistics such as the S-estimators (Davies, 2007), M-estimators (Maronna, 2007), and minimum covariance determinant (MCD) estimators (Baz-Lomba, Harman, Reid, & Thomas, 2017; Seheult, Green, Rousseeuw, & Leroy, 1989) have been developed to detect outliers. Zhu, Shi, Song, Tao, and Tan (2020) proposed an information concentrated variational auto-encoder (IFCVAE) by dividing latent variables into quality-related and unrelated spaces. The other weight-based studies include Spherical PCA (sPCA) (Stanimirova, Daszykowski, & Walczak, 2007), weighted PPCA (WPPCA) (Yuan et al., 2017), divergence-based robust ICA (Chen, Hung, Komori, Huang, & Eguchi, 2013; Mihoko & Eguchi, 2002), a meta gradient descent-based reweighting (Ren, Zeng, & Yang, 2018), and robust Bayesian method (Chen & Ge, 2020).

In conventional neural network training, each data sample is treated equally during parameter learning, regardless of whether it is trustworthy or not. The low-quality data can distort model parameters leading to inferior process monitoring performance. Several methods treat low-quality data by discarding them that can potentially result in the loss of information. We have developed a robust semi-supervised artificial neural network for detecting and isolating faults in process systems. The approach is based on harnessing the data quality feature of a given dataset. The model calculates each sample's quality and reliability based on the Mahalanobis distances and the trusted center of each class of data (Chen & Ge, 2020). The training accounts for reliability and supervisory label to determine parameters of the neural structure. The proposed method improves the accuracy and robustness of a neural network by reducing the influence of unreliable data without sacrificing information by rejecting samples. The subsequent part of the paper is structured in sections representing methods, results and discussion followed by the conclusion.

3.2. Robust neural network

Artificial neural network (ANN) is a nonlinear model inspired by biological neural systems. A standard ANN consists of many processors, also called neurons, which generate a sequence of real-valued activations (Agatonovic-Kustrin & Beresford, 2000; Schmidhuber, 2015). ANNs are universal approximators and can assimilate highly complex relationships between input and output variables precisely (Park & Sandberg, 1991). The learning of detecting process faults is based on pattern recognition properties of neural networks. Fig 3.1 presents the proposed robust neural network model for detecting faults in process systems. It is a multi-layer network comprising an input layer, a hidden layer, and an output layer. The model is different from the simple ANN in terms of considering the quality of samples determined by the model. The network quantifies the quality of a sample by a parallel operation to the forward phase of the training. The quality along with the true and estimated labels computes the loss of the sample. A higher weight is assigned to high-quality data in the training. This is in contrast to the standard neural network where each data is treated equally.
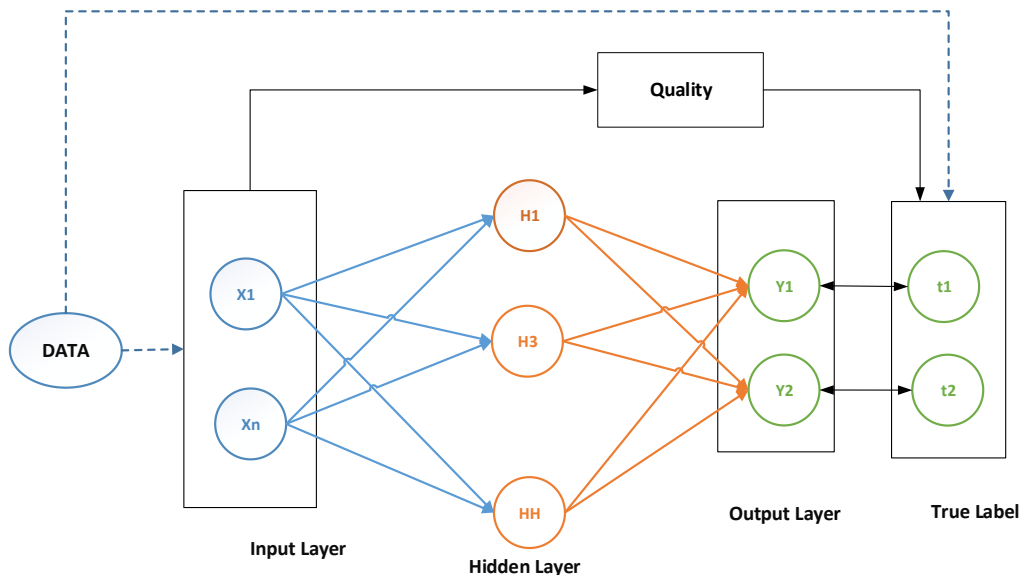
Fig. 3.1: Fault detection using quality-based training of a supervised neural network

The model parameters are iteratively determined using the forward prediction and backward propagation through the network layers. The parameter learning of the network can broadly be summarized under three steps:

Step I: Forward propagation

The forward propagation of the network facilitates predictions based on the input and the network parameters. Eq 3.1-3.4 represent the forward-pass model, where X is the input variables, W_1 and W_2 denote weights and a_1 and a_2 are the activations of the network. We have employed a hyperbolic tangent activation (Eq. 3.2) in the hidden layer to avoid saturation that is observed with the sigmoid function (Glorot & Bengio, 2010). Eq. 3.4 presents the Softmax outputs for yielding an effective classification of faulty and normal scenarios.

$$z_1 = X.W_1 \qquad (3.1)$$

$$a_1 = \tanh z_1 \qquad (3.2)$$

$$z_2 = a_1.W_2 \qquad (3.3)$$

$$a_{2j} = \frac{\exp(z_{2j})}{\sum_k \exp(z_{2k})} \qquad (3.4)$$

Step II: Calculate the cost function

The loss function evaluates the "goodness" of the model predictions. It is a measure of the cost caused by incorrect predictions. We have employed the entropy loss function for training the model. Entropy is positive and tends to zero if the output of all samples matches with their targets. The cross-entropy cost function does not slow down parameter learning which is frequently encountered in quadratic cost (Nelson, 2020). Eq. 3.5 presents the overall loss weighted by the quality of the samples characterizing the stronger influence of higher quality samples.

$$L = -q_n \sum_i t_i \log a_{2i} \qquad (3.5)$$

Step 3: Backpropagation and weight update

The network parameter Θ is determined by backpropagation which is based on the input features, true label, and data quality. The goal of the backpropagation is to optimize the parameter of the network for arbitrarily mapping inputs to outputs. The derivative of the loss function with respect to the network parameter is determined by chain rule (Eq. 3.6-3.13) (Aggarwal & Murty, 2021; Hinton, 1986)-

$$
\begin{aligned}
\frac{\partial L}{\partial W_{2ji}} &= \frac{\partial L}{\partial a_{2i}} \frac{\partial a_{2i}}{\partial z_{2i}} \frac{\partial z_{2i}}{\partial W_{2hi}} \\
&= \frac{\partial L}{\partial z_{2i}} \frac{\partial z_{2i}}{\partial W_{2hi}} \\
&= -q_n \sum_h t_h \frac{\partial \log a_{2i}}{\partial z_{2i}} \frac{\partial z_{2i}}{\partial W_{2hi}} \\
&= -q_n \sum_h \frac{t_i}{a_{2i}} \frac{\partial a_{2i}}{\partial z_{2i}} \frac{\partial z_{2i}}{\partial W_{2hi}} \\
&= -q_n \sum_h \frac{t_i}{a_{2i}} \frac{\partial a_{2i}}{\partial z_{2i}} a_{1h} \\
&= -q_n \left(\sum_{h=i} \frac{t_i}{a_{2i}} \frac{\partial a_{2i}}{\partial z_{2i}} + \sum_{h\neq i} \frac{t_i}{a_{2i}} \frac{\partial a_{2i}}{\partial z_{2i}} \right) a_{1h}
\end{aligned}
\qquad (3.6)
$$

The differential of the Softmax function can be presented by Eq. 3.7 (Arat, 2019).

$$
\begin{aligned}
\frac{\partial a_{2i}}{\partial z_{2i}} &= a_{2i}(1 - a_{2i}) && \forall\, i = h \\
\frac{\partial a_{2i}}{\partial z_{2i}} &= -a_{2h} a_{2i} && \forall\, i \neq j
\end{aligned}
\qquad (3.7)
$$

Substituiting $\frac{\partial a_{2i}}{\partial z_{2i}}$, Eq. 3.6 can be represented as follows-

$$
\frac{\partial L}{\partial W_{2hi}} = -q_n \left\{ \frac{t_i}{a_{2i}} a_{2i}(1 - a_{2i}) - \sum_{h\neq i} \frac{t_j}{a_{2h}} (-a_{2h} a_{2i}) \right\} a_{1h}
$$

$$= -q_n \left(-t_i + t_i a_{2i} + \sum_{h \neq i} t_j a_{2i} \right) a_{1h}$$

$$= -q_n \left(-t_i + a_{2i} \sum_h t_h \right) a_{1h}$$

$$= q_n (t_i - a_{2i}) a_{1h} \qquad (3.8)$$

Where, $\sum_h t_h = 1$, as only one label, will be true. Similarly, $\dfrac{\partial L}{\partial W_{1kh}}$ can be represented as follows-

$$\frac{\partial L}{\partial W_{1kh}} = \frac{\partial L}{\partial a_{1h}} \frac{\partial a_{1h}}{\partial z_{1h}} \frac{\partial z_1 h}{\partial W_1 kh} \qquad (3.9)$$

$$\frac{\partial L}{\partial a_{1h}} = \frac{\partial L}{\partial a_{2i}} \frac{\partial a_{2i}}{\partial z_{2i}} \frac{\partial z_{2i}}{\partial a_{1h}} = q_n(t_i - a_{2i})W_{2hi} \qquad (3.10)$$

$$\frac{\partial a_{1h}}{\partial z_{1h}} = 1 - a_{1h}^2 \qquad (3.11)$$

$$\frac{\partial z_{1h}}{\partial W_{1kh}} = x_k \qquad (3.12)$$

$$\frac{\partial L}{\partial W_{1kh}} = q_n(t_i - a_{2i})W_{2hi}(1 - a_{1h}^2)x_k \qquad (3.13)$$

The gradients adjust the parameters using the *parameter update equations* Eq.3.14-3.15, where η is the *'learning rate'* that controls the influence of the current gradient on the weight update. We have employed a step decay adaptive learning rate with a relatively high learning rate at the beginning that lowers stepwise during the training. We can see that the sample quality impacts weight revision and convergence in estimating the model parameters. The next section presents the procedure for estimating the quality of the sample.

$$W_{1kh} = W_{1kh} - \eta \frac{dL}{dW_{1kh}} \qquad (3.14)$$

$$W_{2hi} = W_{2hi} - \eta \frac{dL}{dW_{2hi}} \qquad (3.15)$$

3.2.1: *Estimation of quality of samples*

Data quality is a broad term characterizing the fitness for the requirements of its intended function. The dimension of data quality includes accuracy, completeness, consistency, integrity, and timeliness (Batini, Cappiello, Francalanci, & Maurino, 2009). A low-quality data can be materialized due to sensor breakdown, process shutdown, malfunctioning of equipment, random fluctuations, incorrect calibration, inconsistent sampling frequencies, and data entry errors due to human factors. Industrial systems encounter various low data-quality problems such as incomplete data, outliers, imbalanced data, poorly labeled data, incorrectly mapped properties, noise, and inconsistency.

We have studied the effect of mislabeled data in the fault detection system by introducing 1%, 5%, and 10% mislabeling in the training and testing samples. The x% mislabeling indicates x% of the training data were assigned incorrect labels. For instance, 1% of the faulty samples were mislabeled as normal in the training of the model. Suppose that X $(X \in R^{NXD})$ is a training dataset where N = Number of samples and D = Number of variables. The first 'i' number of data (1 to i) have label L1, and the remaining (N-i) data have label L2 as shown in Table 3.1. The labels of 'r' samples are misreported (Table 3.2).

The procedure for calculation of the quality and the corresponding weight for the training is as follows:

1. The process data is obtained from the historical database D denoted as $X \in R^{NXD}$ where N (row) and D (column) represent the number of samples and number of process variables, respectively.

2. Calculation of trusted center denoted as X_{trust} on a reliable dataset based on the dependent and independent variables.

Table 3.1. True data samples

Feature					Label
X11	X12	X13	....	X1d	L1
X21	X22	X23	....	X2d	L1
X31	X32	X33	....	X3d	L1
Xi1	Xi2	Xi3	....	Xid	L1
Xi+11	Xi+12	Xi+13	....	Xi+1d	L2
.					
Xi+r1	Xi+r2	Xi+r3	....	Xi+rd	L2
Xi+r+11	Xi+r+12	Xi+r+13	....	Xi+r+1d	L2
X1	Xn2	Xn3	....	Xnd	L2

Table 3.2: Mislabeled samples/data

Feature					Label
X11	X12	X13	....	X1d	L1
X21	X22	X23	....	X2d	L1
X31	X32	X33	....	X3d	L1
Xi1	Xi2	Xi3	....	Xid	L1
Xi+11	Xi+12	Xi+13	....	Xi+1d	L1
.					
Xi+r1	Xi+r2	Xi+r3	....	Xi+rd	L1
Xi+r+11	Xi+r+12	Xi+r+13	....	Xi+r+1d	L2
Xn1	Xn2	Xn3	....	Xnd	L2

Mislabeled data
True label= L2
Mislabeled as L1
label= L1

3. Calculation of Mahalanobis distance: The Mahalanobis distance for an unreliable sample X_n can be calculated by Eq. 3.17, where S is the covariance matrix of process variables evaluated from X_R (Chen & Ge, 2020).

$$d_n = \sqrt{(X_n - X_{trust})^T S^{-1}(X_n - X_{trust})} \qquad (3.17)$$

4. Calculation of the quality of each sample using Eq.(3.18) (Chen & Ge, 2020).

$$q_n = \exp\left(-\eta_d \frac{d_n}{\varphi_d}\right) \qquad (3.18)$$

where φ_d represents the standard deviation of $d = \{d_1, d_2, \ldots \ldots d_n\}$ and η_d is a tuning factor to adjust the calculation sensitivity. The overall quality feature results for unreliable process data can be represented as follows.

$$\mathbf{q} = \{q_1, q_2, \ldots \ldots q_n\} \qquad (3.19)$$

3.2.2: *Fault detection using the robust ANN*

Fig. 3.2 presents the flow diagram of detecting faults using quality-based training of a semi-supervised neural network. A multi-layer feedforward neural network with a hyperbolic tangent activation at hidden layers and Softmax activation at the output layer have been used to classify the normal and faulty data. The neural network structure is based on selecting an effective optimization algorithm and an optimal number of neurons in the hidden layer of the network. The data is normalized before feeding to the model to avoid the scaling effect. The normalized data is divided into training, validation, and testing. The training set creates a model that attempts to fit each data in the training set accurately irrespective of the model complexity. The validation is used to avoid overfitting and compromised generalization of the model. The network structure for fault detection and isolation is determined iteratively by varying optimization algorithms, the number of neurons in the hidden layer, weights, and bises leading to optimal performance. The sample quality plays a significant role in estimating parameters of the fault detection system.
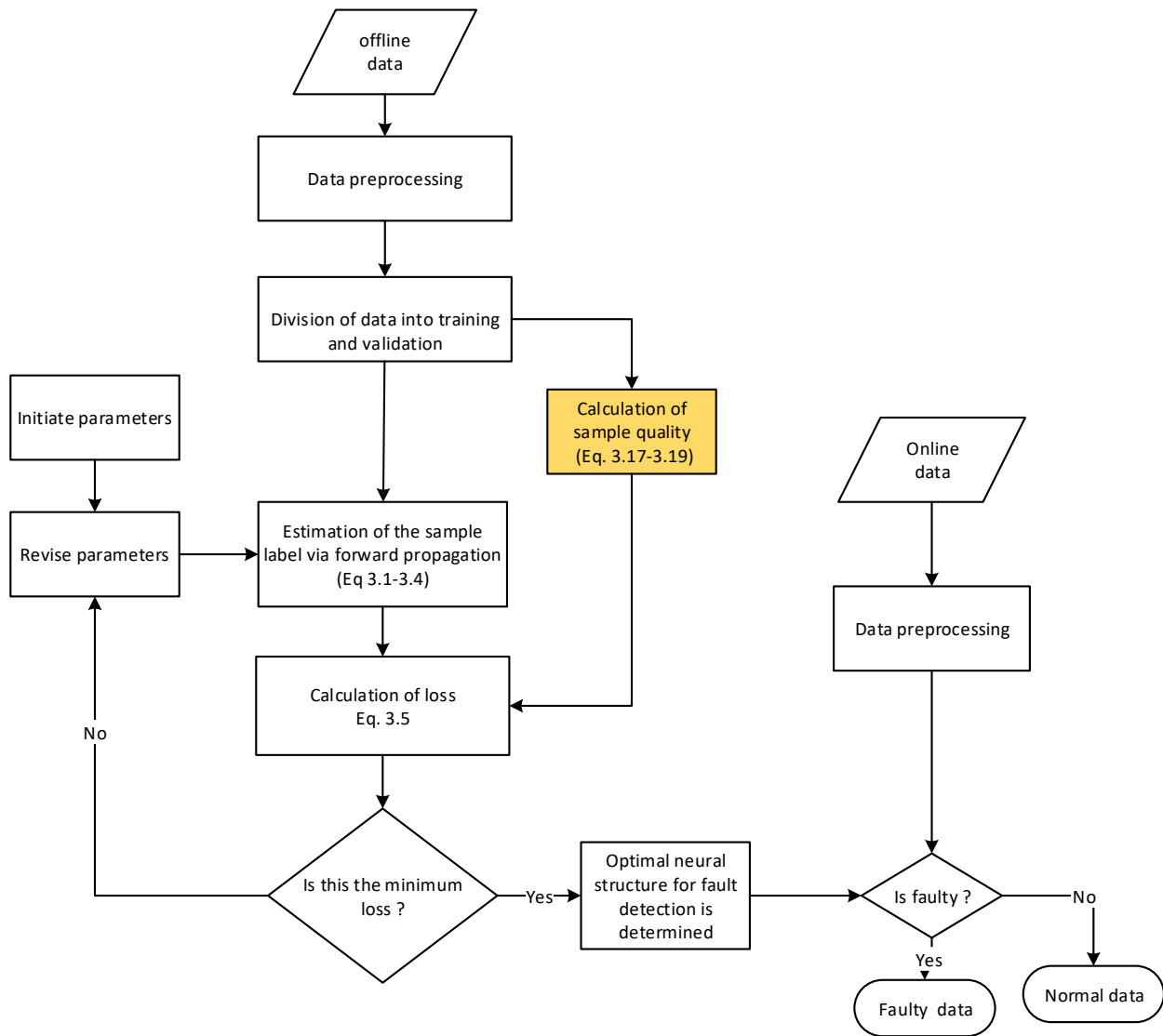
Fig. 3.2: Flowchart of fault detection using quality-based training of a supervised neural network

3.3 Results and Discussion

This section presents the evaluation of the proposed model on detecting faults in two case studies: leak detection in a CSTH (Section 4.3.1) and fault detection of the Tennessee Eastman Process (Section 4.3.2). The following section presents an overview of the metrics employed in the evaluation.

The performance of a classification model is usually assessed on three key parameters; accuracy, fault detection rate (FDR), and false alarm rate (FAR). Accuracy represents the overall effectiveness of a classifier and can be determined by the fraction of correct prediction (Eq. 3.20). The FDR announces the proportion of the faults rightly detected by a classifier (Eq. 3.21) whereas FAR indicates the fraction of the alarms turned out false (Eq. 3.22). A good classification system should have higher accuracy and FDR as well as a lower value of the FAR (Zhao et al., 2017; Alauddin, Khan, Imtiaz, & Ahmed, 2020).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \qquad (3.20)$$

$$FDR = \frac{Fault\ detected}{Total\ number\ of\ faults} = \frac{TP}{TP+FP} \qquad (3.21)$$

$$FAR = \frac{False\ alarm}{Total\ number\ of\ alarms} = \frac{FN}{TP+FN} \qquad (3.22)$$

Where, TP= true positive; fault signal correctly diagnosed as a fault.

TN= true negative; normal signal correctly diagnosed as normal

FP= false positive; fault signal incorrectly identified as normal.

FN= false negative; normal signal incorrectly identified as a fault

Fault detection rate, false alarm rate, and accuracy are commonly used metrics for evaluating classification models. However, they do not alone can provide a holistic view of the classifier's efficacy and can lead to deceptive outcomes especially for the imbalanced data. For instance, a fault detection system with a 100% detection rate (FDR=1) is good with respect to detecting faults as it is not missing detecting abnormal signals. Nonetheless, the system might be biased towards detecting faults and can predict a normal signal as faulty. This can result in excessive false alarms. Similarly, a detecting system with negligible FAR might be prone to missing detecting faulty scenarios. Accuracy could give a balanced estimate of classifying faults as faults and normal as

normal. However, its results could be misleading in the case of imbalanced samples. For example, a classifier with a 9:1 ratio of normal to faulty samples in a binary classification could register 90% accuracy by categorizing all signals as normal. In this case, the classification system is not detecting any fault despite a higher accuracy value.

We have devised a new metric, the classification index CI ($0 \leq CI \leq 2$), for evaluating binary classifiers considering accuracy, fault detection, and missed detection of abnormal situations. Eq. 3.23 presents the classification index where, w_{FDR} and w_{FAR} denotes respective importance parameters for FDR and FAR. Thus an FDD system with CI=2 guarantees effective detection of faults as the fault and normal as normal signals.

$$Classification\ Index\ (CI) = Accuracy + w_{FDR} * FDR - w_{FAR} * FAR \qquad (3.23)$$

For an unconstrained classifier system, $w_{FDR} = w_{FAR} = 1$, leading to the classification index, $(CI) = Accuracy + FDR - FAR$. Higher the CI better will be the classifier. The CI is a bounded function with values that lies in [-1, 2]. The value of the classification index is equal to -1 when Accuracy=0, FDR=0, and FAR=1 and equal to 2 for Accuracy=1, FDR=1, and FAR=0. A higher CI value is desirable for efficient classification.

3.3.1 *Leak detection in a continuous stirred tank heater (CSTH)*

The continuous stirred tank heater is a widely used unit operation of chemical process systems. Fig. 3.3 presents a CSTH model developed at the University of Alberta (Thornhill, Patwardhan, & Shah, 2008). The temperature of water in the CSTH is assumed to be uniform throughout the vessel. The state of the output variables (water level, flow rate, and temperature) are recorded for the distinct openings of the inlet valves.

The proposed robust neural network model has been examined on the CSTH model for detecting leakage in the vessel. The data were divided in to training and test samples with a 70: 30 ratio. The robustness of the model was tested by introducing 1%, 5%, and 10% mislabeling in the training and testing samples. The 1% mislabeling indicates 1% of the training data were assigned incorrect labels. The performances of the proposed robust ANNs were compared against standard ANN in terms of fault detection rate (FDR) and a classification index (CI) that account for accuracy, fault detection rate, and false alarm rate on test samples. Fig. 3.4 presents the FDR of the proposed quality-based trained ANN and the standard ANN for 0%, 1%, 5%, and 10% mislabeled data. The 0% mislabeled data represents a legitimate data sample i.e., no normal data has been labeled ad faulty and vice-versa. The FDR of the robust ANN is higher than that of the standard ANN for the same initialization of the network parameters. It can also be noticed that the robust network performs better on legitimate data as well.
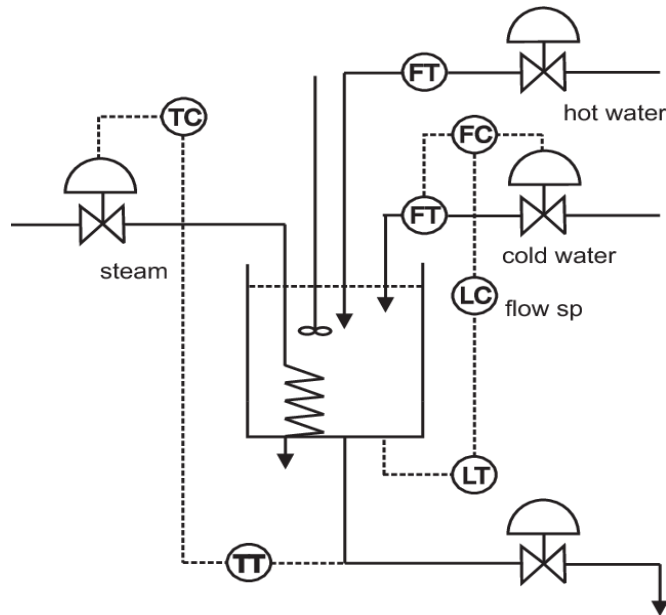
Fig. 3.3: Schematic of a continuous stirred tank heater model (Thornhill, Patwardhan, & Shah, 2008)

77
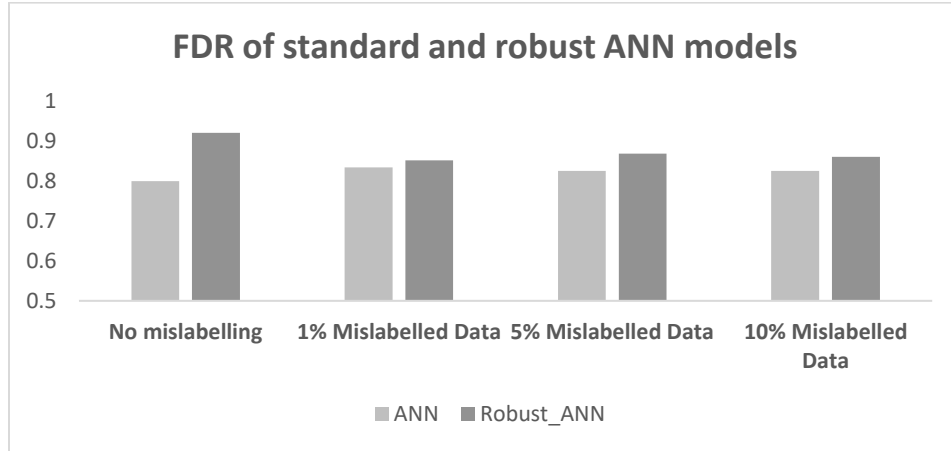
Fig. 3.4: Fault detection rate of leak detection of a CSTH using the proposed Robust ANNs and standard ANN

The fault detection rate alone does not guarantee the goodness of a fault detection model. The threshold of the detecting system might be biased towards detecting abnormalities resulting in a higher false alarm rate (FAR). We have examined the performance of the FDD system using a receiver operating characteristic (ROC) curve. The ROC aids in visualizing the trade-off between FDR and FAR with optimal classification. Fig. 3.5 presents the receiver operating characteristic curves of the fault detecting models based on standard ANN and the proposed robust ANN model. Fig. 5A-C presents the corresponding ROC curves for the 0%, 1%, 5%, and 10% mislabeled data. The random model represented by the dotted diagonal in the ROC curve is a random classifier that can not differentiate the two classes at any threshold. A detection system can distinguish the normal and faulty classes effectively if its ROC curve is farther away from the diagonal. The area under the ROC curve (AUC) represents a degree of separability. The higher the AUC, the better the model at predicting distinct classes. The AUC is classification-threshold-invariant, i.e. it assesses the performance of the model irrespective of the classification threshold.
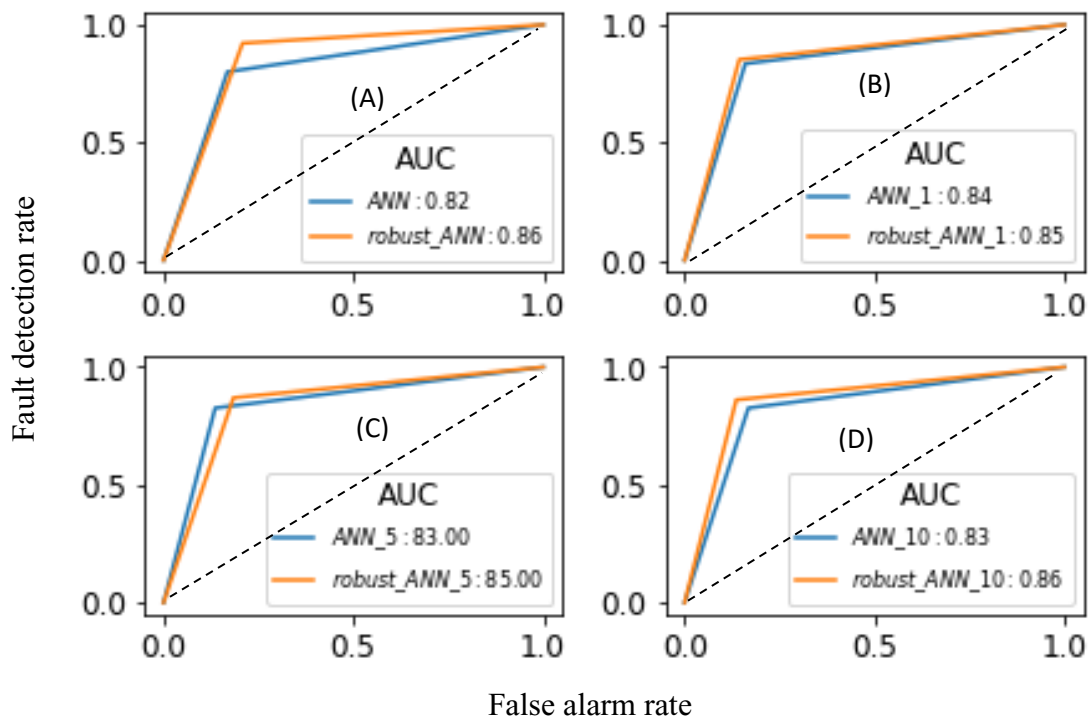
Fig. 3.5: The comparative ROC curves of the Robust ANNs and simple ANN on for leak detection of CSTH: (A) legitimate data, (B) 1% mislabeled data, (C), 5% mislabeled data, and (D) 10% mislabeled data

The models have also been examined based on the classification indices($CI = Accuracy + FDR - FAR$). Here again, the proposed robust ANN model scores well compared to the standard ANN. Thus, the percentage improvements in CIs of the proposed robust model compared to the standard ANN were recorded as 9.5%, 2.8%, 1.9%, and 6.3% on 0%, 1%, 5%, and 10% mislabeled data respectively.

The quality-based robust ANN model performs well for leak detection in a CSTH system. The next section examines its effectiveness for detecting faults of the Tennessee Eastman process.

3.3.2 *Fault detection in the Tennessee Eastman process*

The Tennessee Eastman process is a complex nonlinear simulated process that has been used in many chemometrics studies for plant-wide control, process monitoring, and fault detection and

diagnosis. It comprises five major operating units: an exothermic two-phase reactor, a product condenser, a vapor-liquid flash separator, a recycle compressor, and a reboiled product stripper as shown in Fig. 3.6 (Downs & Vogel, 1993). The TE process constitutes 41 (XMEAS-1 to XMEAS-41) measured and 12 (XMV-1 to XMV-12) manipulated variables. The measured variables include 22 continuous process variables and 19 variables for composition measurements. Twenty distinct known and unknown faults (IDV-1 to IDV20) have been studied by many researchers (Ajagekar & You, 2020; Alauddin et al., 2020; Amin, Khan, Imtiaz, & Ahmed, 2019; Ge & Song, 2007; Tang et al., 2018; Wu & Zhao, 2020; Xie, Yang, Li, & Ji, 2019; Jianbo Yu & Yan, 2019; Zhang, Guo, & Li, 2020)
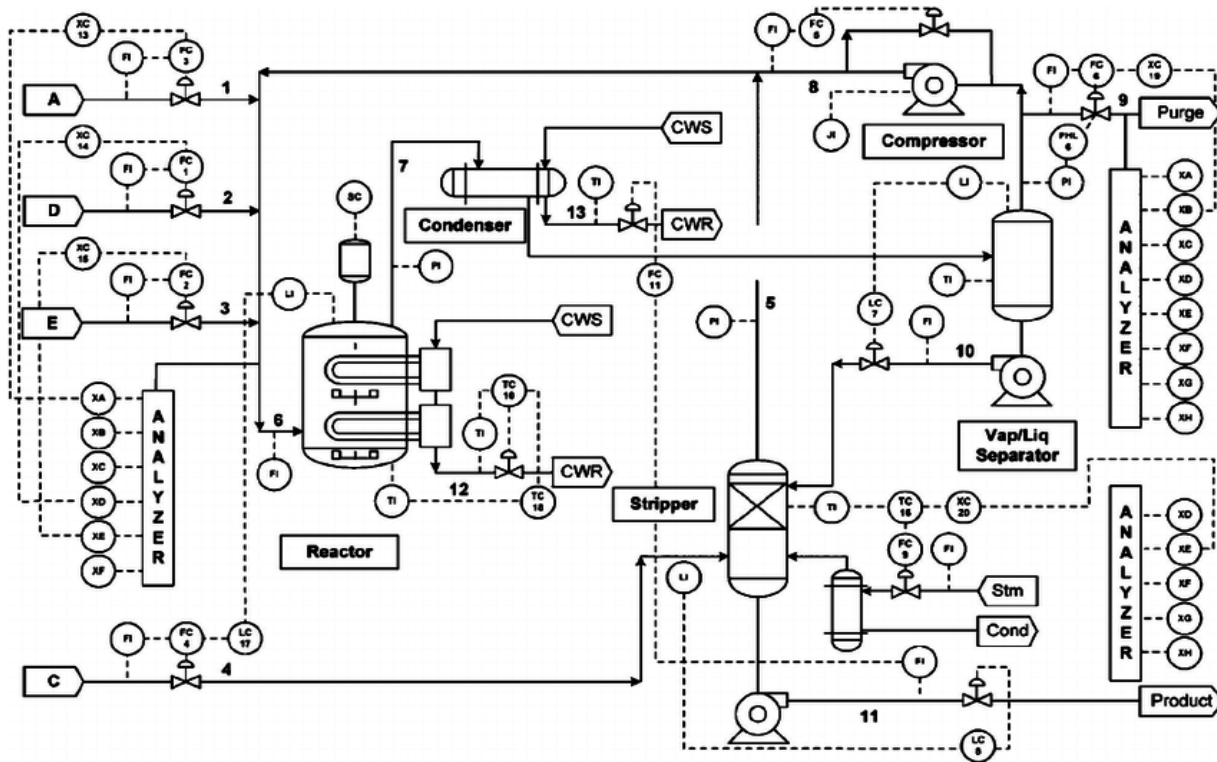
Fig 3. 6: Process flow diagram of the Tennessee Eastman process (Downs & Vogel, 1993)

The proposed robust ANN model has been examined on the Tennessee Eastman process for detecting step fault (IDV-1) and sticking fault (IDV-14).

3.3.2.1 *Fault detection of IDV-1 (step fault) of the TE process*

Fig. 3.7 presents the fault detection rate (FDR) of the proposed robust ANN and simple ANN at 0%, 1%, 5%, and 10% mislabeled data. The data were divided in to training and test samples with a 70: 30 ratio. The models were evaluated on test samples. The FDR of the Robust ANNs is higher than the standard ANN. For instance, FDRs of the robust ANN and standard ANN (in bracket) were found as: 0.94 (0.92), 0.93 (0.93), 0.93 (0.88) and 0.86 (0.78) on respective 0%, 1%, 5%, and 10% mislabeled data. It can also be deduced that the FDR of the standard ANN decreases drastically with increasing mislabeling. Here, a marginal improvement of the FDR of the simple ANN at 1% mislabeling might be due to the random pick up of the smaller number of mislabeled data in the test samples of the ANN compared to that for the robust ANN. Had the test samples be exactly the same, the robust ANN would score better. This can be justified by the accuracy of the models presented in Fig. 3.8. The accuracy, which is denoted by the AUC of the ROC curve, of the proposed robust ANNs is higher than the corresponding simple ANNs for all categories of samples (Fig. 3.8).

The classification indices of the robust ANN and standard ANN (in bracket) were found as: 1.91 (1.79), 1.87 (1.85), 1.88 (1.82) and 1.79 (1.66) on respective 0%, 1%, 5%, and 10% mislabeled data. Thus, the CIs of the robust ANN were improved significantly on all test conditions.
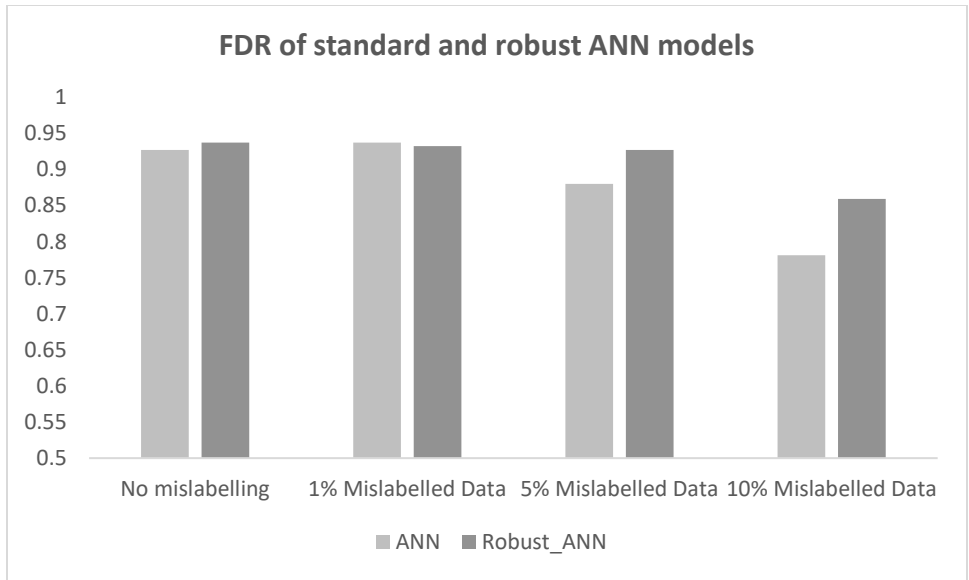
Fig. 3.7: Fault detection rate of the TE process for detecting step fault (IDV-1) using the proposed robust ANNs and standard ANNs
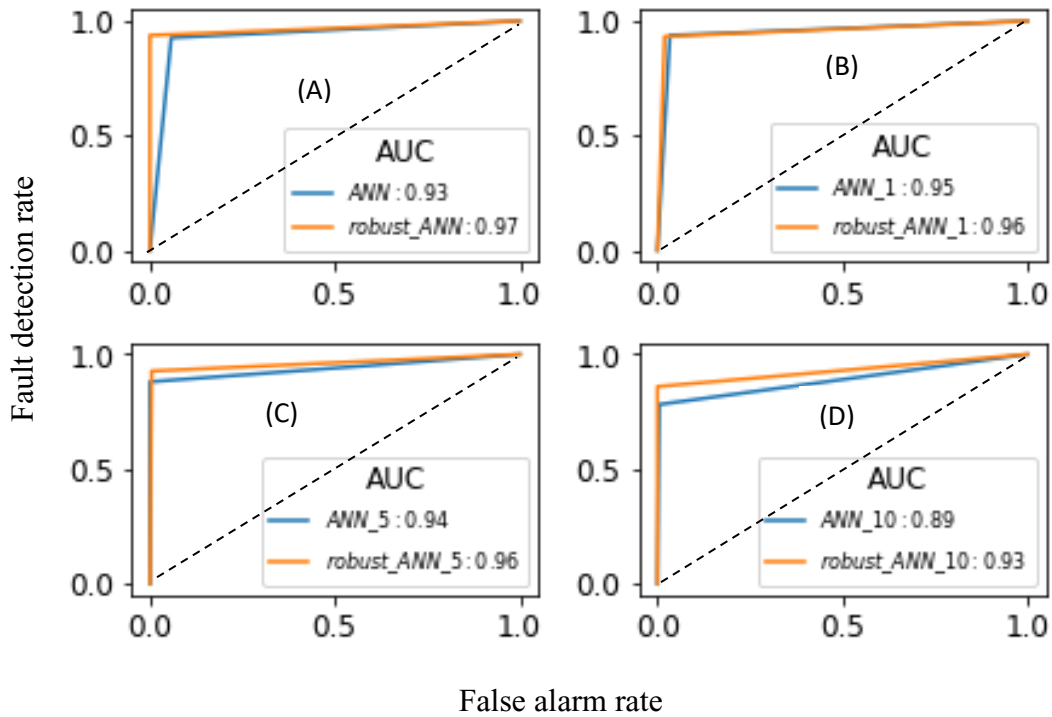
Fig. 3.8: Comparative ROC curves of the robust ANNs and standard ANNs for detecting the step fault (IDV-1) of TE process: (A) legitimate data, (B) 1% mislabeled data, (C), 5% mislabeled data, and (D) 10% mislabeled data

3.3.2.2 *Fault detection of IDV-14 (sticking fault) of the TE process*

The sticking fault of the reactor cooling water valve (IDV-14) increases reactor temperature (XMEAS-9), separator temperature (XMEAS-11), and condenser cooling water outlet temperature (XMEAS-22). The elevated temperature initiates the thermal runaway of the reactor leading to failure.

The comparative performance of the test algorithms for the detection of IDV-14 fault of the TE process has been presented in Table 3.3. We did not see any significant improvement in FDR; however, there is marked gain in the classification index of the robust ANN over the standard ANN. This difference widens with the amount of mislabeled data. The overall effectiveness is reported by the proposed classification index. The higher CI values ($CI = Accuracy + FDR - FAR$) of the proposed algorithms speak for the superior efficacy of the proposed robust neural network-based fault detection systems.

Table 3.3: Comparative performance of the robust ANN and standard ANN on IDV-14 of the TE process

		1% mislabeled data	*5% mislabeled data*	*10% mislabeled data*
Fault	*Simple ANN*	0.93	0.93	0.86
detection rate	***Robust ANN***	**0.93**	**0.94**	**0.93**
Classification	*Simple ANN*	1.71	1.61	1.39
index	***Robust ANN***	**1.73**	**1.69**	**1.61**

3.4 Conclusions

This work presents a novel approach, the quality-based training (QbT) of the supervised neural networks, for handling the mislabeled and/or low-quality data. The quality of each data sample was calculated by the Mahalanobis distance of a sample from a trusted center. The model was evaluated for two case studies; a constant stirred tank heater and the Tennessee Eastman chemical process. The effectiveness of detecting faults of the proposed quality-based trained robust neural networks was compared with the conventional artificial neural network-based model for the same random numbers and related hyper-parameters. The proposed robust ANN registered a higher fault detection rate and classification indices on distinct mislabeled data. The percentage improvements in accuracy in detecting IDV-1 fault of the TE process were recorded as 1%, and 4.5 % on respective 1%, and 10% mislabeled data. In terms of the proposed classification indices (CI), the improvements amount to 1.1%, and 7.5% on corresponding test samples. The proposed method also yielded improved results in detecting faults on legitimate samples i.e., samples with no mislabeling. Thus, respective gains in accuracy and the CI in detecting faults on legitimate samples were 4.30% and 6.33%. Similar leverages were noticed in detecting the sticking fault (IDV-14) of the TE process and leak in the CSTH system using the proposed robust ANN.

We studied the effect of the random mislabeled data. The proposed model can be studied for incorporating causality and structural relations. The proposed quality-based training is effective in detecting faults in presence of the low-quality data.

Acknowledgment

List of symbols and abbreviations

Symbols	Meanings
ANN	artificial neural network
a_1	activation for the hidden layer
a_{1h}	h^{th} element of the hidden activation layer
a_2	activation for the output layer
a_{2j}	j^{th} element of the output activation layer
AUC	area under the receiver operating characteristic curve
CART	classification and regression tree
CBR	case-based reasoning
CI	classification index
d_n	Mahalanobis distance of nth sample wrt the trust centre
FAR	false alarm rate
FDD	fault detection and diagnosis
FDR	fault detection rate
FN	false negative
FP	false positive
GMM	Gaussian mixture model
ICA	independent component analysis
IDV-1 to 20	faults of the simulated TE Process
IFCVAE	information concentrated variational auto-encoder
L	loss
L1	label 1
L2	label 2
PCA	principle component analysis
PLS	partial least square
QbT	quality-based training
q_n	quality of the nth sample
RNN	recurrent neural networks
ROC	receiver operating characteristic
S	covariance matrix of process variables
sPCA	spherical principal component analysis
SVM	support vector machine
t_i	i^{th} label of the target
TN	true negative
TP	true positive
W_1	weight matrix for the hidden layer
W_2	weight matrix for the output layer
W_{2ji}	weight of the jth neuron to i^{th} output

w_{FAR}	importance parameters for the false alarm rate
w_{FDR}	importance parameters for the fault detection rate
X	training dataset
x_k	k^{th} column of the input matrix
XMEAS	input variable in the TE Process
Xnd	elements corresponding to n^{th} row and d^{th} column of training sample
X_{trust}	trusted centre of the reliable dataset
z_1	hidden layer vector
z_{1k}	k^{th} element of the hidden layer
z_2	output layer vector
z_{2k}	k^{th} element of the output layer
φ_d	standard deviation of d samples
η	learning rate
η_d	tuning factor to adjust the calculation sensitivity
$\mathbf{q}$	quality vector of the data

References

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Aggarwal, M., & Murty, M. N. (2021). *Deep Learning. SpringerBriefs in Applied Sciences and Technology*. https://doi.org/10.1007/978-981-33-4022-0_3

Ajagekar, A., & You, F. (2020). Quantum computing assisted deep learning for fault detection and diagnosis in industrial process systems. *Computers and Chemical Engineering*. 143, 107119. https://doi.org/10.1016/j.compchemeng.2020.107119

Alauddin, M., Khan, F., Imtiaz, S., & Ahmed, S. (2018). A Bibliometric Review and Analysis of Data-Driven Fault Detection and Diagnosis Methods for Process Systems. *Industrial and Engineering Chemistry Research*, *57*(32), 10719–10735. review-article. https://doi.org/10.1021/acs.iecr.8b00936

Alauddin, M., Khan, F., Imtiaz, S., & Ahmed, S. (2020). A variable mosquito flying optimization-based hybrid artificial neural network model for the alarm tuning of process fault detection systems. *Process Safety Progress*, *39*(S1), e12122 . https://doi.org/10.1002/prs.12122

Amin, M. T., Khan, F., Imtiaz, S., & Ahmed, S. (2019). Robust Process Monitoring Methodology for Detection and Diagnosis of Unobservable Faults. *Industrial and Engineering Chemistry Research*, *58*(41), 19149–19165. https://doi.org/10.1021/acs.iecr.9b03406

Arat, M., M. (2019), Derivation of Softmax Function, https://mmuratarat.github.io/2019-01-27/derivation-of-softmax-function

Arunthavanathan, R., Khan, F., Ahmed, S., & Imtiaz, S. (2021). An analysis of process fault diagnosis methods from safety perspectives. *Computers & Chemical Engineering*, *145*, 107197. https://doi.org/10.1016/j.compchemeng.2020.107197

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41(3), 1-52*. https://doi.org/10.1145/1541880.1541883

Baz-Lomba, J. A., Harman, C., Reid, M., & Thomas, K. V. (2017). Passive sampling of wastewater as a tool for the long-term monitoring of community exposure: Illicit and prescription drug trends as a proof of concept. *Water Research*, *121*, 221–230. https://doi.org/10.1016/j.watres.2017.05.041

Berners-Lee, T., & Shadbolt, N. (2011). There's gold to be mined from all our data. *Times*. http://www.thetimes.co.uk/tto/...272618.ece

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Cai, L., & Tian, X. (2014). A new fault detection method for non-Gaussian process based on robust independent component analysis. *Process Safety and Environmental Protection*, *92*(6), 645–658. https://doi.org/10.1016/j.psep.2013.11.003

Cai, L., Tian, X., & Chen, S. (2014). A process monitoring method based on noisy independent component analysis. *Neurocomputing*, *127*, 231–246. https://doi.org/10.1016/j.neucom.2013.07.029

Chen, G., & Ge, Z. (2020). Robust Bayesian networks for low-quality data modeling and process monitoring applications. *Control Engineering Practice*, *97*(June 2019), 104344. https://doi.org/10.1016/j.conengprac.2020.104344

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20, 273-297*. https://doi.org/10.1023/A:1022627411411

Cui, Q., & Li, S. (2020). Process monitoring method based on correlation variable classification and vine copula. *Canadian Journal of Chemical Engineering*, *98*(6), 1411–1428. https://doi.org/10.1002/cjce.23702

Dai, X., & Gao, Z. (2013). From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis. *IEEE Transactions on Industrial Informatics*, *9*(4), 2226–2238. https://doi.org/10.1109/TII.2013.2243743

Davies, P. L. (2007). Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *The Annals of Statistics*, *15*(3), 1269–1292. https://doi.org/10.1214/aos/1176350505

Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers and Chemical Engineering*, *17*(3), 245–255. https://doi.org/10.1016/0098-1354(93)80018-I

Gao, Z., Cecati, C., & Ding, S. X. (2015). A survey of fault diagnosis and fault-tolerant techniques-part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, *62*(6), 3757–3767. https://doi.org/10.1109/TIE.2015.2417501

Ge, Z. (2017). Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemometrics and Intelligent Laboratory Systems*, *171*, 16–25. https://doi.org/10.1016/j.chemolab.2017.09.021

Ge, Z., & Song, Z. (2007). Process monitoring based on independent Component Analysis-Principal Component Analysis (ICA-PCA) and similarity factors. *Industrial and Engineering Chemistry Research, 46(7),* 2054-2063. https://doi.org/10.1021/ie061083g

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Journal of Machine Learning Research* (Vol. 9, pp. 249–256).

Hinton, G. (1986). How the backpropagation algorithm works Warm up : a fast matrix-based approach to computing the output. In *Neural Networks and Machine Learning* (pp. 1–25).

Imtiaz, S. A., & Shah, S. L. (2008). Treatment of missing values in process data analysis. *Canadian Journal of Chemical Engineering*, *86*(5), 838–858. https://doi.org/10.1002/cjce.20099

Jin, X., Wang, S., Huang, B., & Forbes, F. (2012). Multiple model based LPV soft sensor development with irregular/missing process output measurement. *Control Engineering Practice*, *20*(2), 165–172. https://doi.org/10.1016/j.conengprac.2011.10.007

Jossen, S. (2017). The world's most valuable resource is no longer oil, but data. *The Economist*, *423*, 5–8. Retrieved from https://www.economist.com/leaders/2017/05/06/the-worlds-most-

valuable-resource-is-no-longer-oil-but-data

Kaspar, M. H., & Harmon Ray, W. (1993). Dynamic PLS modelling for process control. *Chemical Engineering Science*, *48*(20), 3447–3461. https://doi.org/10.1016/0009-2509(93)85001-6

Khatibisepehr, S., & Huang, B. (2008). Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Industrial and Engineering Chemistry Research*, *47*(22), 8713–8723. https://doi.org/10.1021/ie800386v

Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, *28*(1), 3–21. https://doi.org/10.1016/0169-7439(95)80036-9

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote Sensing*, *67*(10), 1137–1142.

Li, G., Qin, S. J., & Yuan, T. (2016). Data-driven root cause diagnosis of faults in process industries. *Chemometrics and Intelligent Laboratory Systems*, *159*, 1–11. https://doi.org/10.1016/j.chemolab.2016.09.006

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, *231*, 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114

Maronna, R. A. (1976). Robust M-Estimators of Multivariate Location and Scatter. *The Annals of Statistics*, *4*(1), 51–67. https://doi.org/10.1214/aos/1176343347

Nelson, M., The cross-entropy cost function. (2020, December 1). Retrieved June 20, 2021, from https://eng.libretexts.org/@go/page/3752

Park, J., & Sandberg, I. W. (1991). Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, *3*(2), 246–257. https://doi.org/10.1162/neco.1991.3.2.246

Qin, S. J. (1998). Recursive PLS algorithms for adaptive data modeling. *Computers and Chemical Engineering*, *22*(4–5), 503–514. https://doi.org/10.1016/s0098-1354(97)00262-7

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control, 36(2), 220-234*. https://doi.org/10.1016/j.arcontrol.2012.09.004

Rashid, M. M., & Yu, J. (2012). Hidden markov model based adaptive independent component analysis approach for complex chemical process monitoring and fault detection. *Industrial*

and Engineering Chemistry Research, *51*(15), 5506–5514. https://doi.org/10.1021/ie300203u

Ren M, Zeng W, Yang B, Urtasun R. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning 2018* Jul 3 (pp. 4334-4343). *PMLR.*

Ren, X., Zhu, K., Cai, T., & Li, S. (2017). Fault Detection and Diagnosis for Nonlinear and Non-Gaussian Processes Based on Copula Subspace Division. *Industrial and Engineering Chemistry Research*, *56*(40), 11545–11564. https://doi.org/10.1021/acs.iecr.7b02419

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Seheult, A. H., Green, P. J., Rousseeuw, P. J., & Leroy, A. M. (1989). Robust Regression and Outlier Detection. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *152*(1), 133. https://doi.org/10.2307/2982847

Smyth, P. (1994). Hidden Markov models for fault detection in dynamic systems. *Pattern Recognition*, *27*(1), 149–164. https://doi.org/10.1016/0031-3203(94)90024-8

Stefatos, G., & Hamza, A. Ben. (2010). Dynamic independent component analysis approach for fault detection and diagnosis. *Expert Systems with Applications*, *37*(12), 8606–8617. https://doi.org/10.1016/j.eswa.2010.06.101

Tang, P., Peng, K., Zhang, K., Chen, Z., Yang, X., & Li, L. (2018). A Deep Belief Network-based Fault Detection Method for Nonlinear Processes. IFAC-PapersOnline, 51(24),9-14 https://doi.org/10.1016/j.ifacol.2018.09.522

Thissen, U., Swierenga, H., De Weijer, A. P., Wehrens, R., Melssen, W. J., & Buydens, L. M. C. (2005). Multivariate statistical process control using mixture modelling. *Journal of Chemometrics*, *19*(1), 23–31. https://doi.org/10.1002/cem.903

Thornhill, N. F., Patwardhan, S. C., & Shah, S. L. (2008). A continuous stirred tank heater simulation model with applications. *Journal of Process Control*, *18*(3), 347–360. https://doi.org/10.1016/j.jprocont.2007.07.006

Tidriri, K., Chatti, N., Verron, S., & Tiplica, T. (2016). Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control, 42, 63-81.* https://doi.org/10.1016/j.arcontrol.2016.09.008

Widodo, A., & Yang, B. S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *21*(6), 2560–2574. https://doi.org/10.1016/j.ymssp.2006.12.007

Wu, H., & Zhao, J. (2020). Fault detection and diagnosis based on transfer learning for multimode chemical processes. *Computers and Chemical Engineering, 135, 106731*. https://doi.org/10.1016/j.compchemeng.2020.106731

Xie, Z., Yang, X., Li, A., & Ji, Z. (2019). Fault diagnosis in industrial chemical processes using optimal probabilistic neural network. *Canadian Journal of Chemical Engineering*, *97*(9), 2453–2464. https://doi.org/10.1002/cjce.23491

Yi, X., Liu, F., Liu, J., & Jin, H. (2014). Building a network highway for big data: Architecture and challenges. *IEEE Network*, *28*(4), 5–13. https://doi.org/10.1109/MNET.2014.6863125

Yin, S., Ding, S. X., Haghani, A., Hao, H., & Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control*, *22*(9), 1567–1581. https://doi.org/10.1016/j.jprocont.2012.06.009

Yin, S., Ding, S. X., Xie, X., & Luo, H. (2014). A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics, 61(11), 6418-6428*. https://doi.org/10.1109/TIE.2014.2301773

Yu, H., Khan, F., & Garaniya, V. (2015). A probabilistic multivariate method for fault diagnosis of industrial processes. *Chemical Engineering Research and Design*, *104*(3), 306–318. https://doi.org/10.1016/j.cherd.2015.08.026

Yu, Jianbo, & Yan, X. (2019). Active features extracted by deep belief network for process monitoring. *ISA Transactions*, *84*, 247–261. https://doi.org/10.1016/j.isatra.2018.10.011

Yu, Jie. (2012). Multiway discrete hidden Markov model-based approach for dynamic batch process monitoring and fault classification. *AIChE Journal*, *58*(9), 2714–2725. https://doi.org/10.1002/aic.12794

Yu, Jie. (2013). A new fault diagnosis method of multimode processes using Bayesian inference based Gaussian mixture contribution decomposition. *Engineering Applications of Artificial Intelligence*, *26*(1), 456–466. https://doi.org/10.1016/j.engappai.2012.09.003

Yu, Jie, & Qin, S. J. (2008). Multimode process monitoring with bayesian inference-based finite Gaussian mixture models. *AIChE Journal*, *54*(7), 1811–1829.

https://doi.org/10.1002/aic.11515

Yu, Jie, & Qin, S. J. (2009). Multiway gaussian mixture model based multiphase batch process monitoring. *Industrial and Engineering Chemistry Research*, *48*(18), 8585–8594. https://doi.org/10.1021/ie900479g

Yuan, X., Wan, Y., Yan, C., Ge, Z., Song, Z., & Gui, W. (2017). Weighted linear dynamic system for feature representation and soft sensor application in nonlinear dynamic industrial processes. *IEEE Transactions on Industrial Electronics*, *65*(2), 1508–1517. https://doi.org/10.1109/TIE.2017.2733443

Zhang, C., Guo, Q., & Li, Y. (2020). Fault detection in the Tennessee eastman benchmark process using principal component difference based on k-nearest neighbors. *IEEE Access, 8, 49999-50009*. https://doi.org/10.1109/ACCESS.2020.2977421

Zhang, S., Bi, K., & Qiu, T. (2020). Bidirectional Recurrent Neural Network-Based Chemical Process Fault Diagnosis. *Industrial and Engineering Chemistry Research*, *59*(2), 824–834. https://doi.org/10.1021/acs.iecr.9b05885

Zhang, Y., & Qin, S. J. (2007). Fault detection of nonlinear processes using multiway kernel independent component analysis. *Industrial and Engineering Chemistry Research*, *46*(23), 7780–7787. https://doi.org/10.1021/ie070381q

Zhang, Y., & Zhang, Y. (2010). Fault detection of non-Gaussian processes based on modified independent component analysis. *Chemical Engineering Science*, *65*(16), 4630–4639. https://doi.org/10.1016/j.ces.2010.05.010

Zhao, H., Sun, S., & Jin, B. (2018). Sequential Fault Diagnosis Based on LSTM Neural Network. *IEEE Access*, *6*, 12929–12939. https://doi.org/10.1109/ACCESS.2018.2794765

Zhao, Hui, Liu, J., Dong, W., Sun, X., & Ji, Y. (2017). An improved case-based reasoning method and its application on fault diagnosis of Tennessee Eastman process. *Neurocomputing*, *249*, 266–276. https://doi.org/10.1016/j.neucom.2017.04.022

Zhou, D., Li, G., & Qin, S. J. (2010). Total projection to latent structures for process monitoring. *AIChE Journal*, *56*(1), 168–178. https://doi.org/10.1002/aic.11977

Zhu, J., Shi, H., Song, B., Tao, Y., & Tan, S. (2020). Information concentrated variational auto-encoder for quality-related nonlinear process monitoring. *Journal of Process Control*, *94*, 12–25. https://doi.org/10.1016/j.jprocont.2020.08.002

Chapter 4

A primer on integrating process dynamics in data-driven models of chemical processing systems

Preface: Chemical engineering discipline has well-established expertise in science-based frameworks in distinctive domains including materials and energy balances, chemical kinetics, transport phenomena (heat, mass, momentum), fluid mechanics, process control, and thermodynamics. The objective of this chapter is to harness this mechanistic knowledge in developing data-driven models for the safety of process systems. This chapter presents a framework for integrating process dynamics and expert knowledge to enhance robustness of data-driven models. This work can be mapped to the thesis's sub-objective "development of robust data-driven models for the safety of process systems".

Bibliographic citation and authorship statement: I (Md Alauddin) was responsible for the problem formulation, conceptualization, development of codes, formal analysis, writing, and revising the manuscript. The bibliographic citation and contributions of the authors and co-author are presented as follows.

Alauddin, M., Khan, F., Imtiaz, S., Ahmad, S., & Amyotte, P., A primer on integrating process dynamics in data-driven models of chemical processing systems *(draft under committee review)*

> *Md Alauddin: Formal Analysis, Methodology, Software; Investigation, Validation; Writing - Original Draft; Writing - Review & Editing*
>
> *Faisal Khan: Conceptualization, Methodology, Writing - Review & Editing; Supervision; Project administration; Funding acquisition*
>
> *Syed Imtiaz: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition*
>
> *Salim Ahmed: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition*
>
> *Paul Amyotte: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition*

Abstract: Data-driven models require high fidelity data of sufficient quantity and granularity. This is challenging in complex chemical processing systems due to frequent sensor breakdown, process shutdown, malfunctioning of equipment, random fluctuations, miscalibration, inconsistent sampling frequencies, and data entry errors. Thus many models scoring well on the training data fails on the real-time data of industrial systems. This work presents a process dynamics-guided deep neural network model to improve the generalization of the model. We have added an additional layer in the neural network architecture to incorporate process dynamics such as material and energy balance equations, universal laws, standard correlations, and field knowledge. We evaluated the proposed models on regression and classification tasks related to processing systems representing steady-state and transient behavior. The results were compared against a standard neural network. The proposed model yielded improved generalization ability on the unseen data. It also produced improved results on models determined by reduced sample-sized data. The proposed process dynamics-guided neural network can be employed as a robust model for handling generalization issues of data-driven methods in processing systems.

4.1   Introduction

4.1.1 Motivation:

The advancement in big data and computations has prompted many industries including process industries to re-examine their traditional roles for design, control, and maintenance. Many data-based tools based on sensitivity analysis and uncertainty quantification have been devised for effective decision making enabling resilient processes and smarter businesses. The state-of-the-art data-driven self-optimizing operations, online monitoring, and control-based simulations of Industy4.0 can reduce potential process upsets and unplanned shutdowns minimizing operational

cost, time, and resources. The data-based models can contribute significantly to making the process industry more efficient, much safer, and environmentally friendlier.

Data-driven methods also known as empirical models are generic and can therefore be adapted to almost any physical or psychological environment (Jack et al., 2018). The data-based models, especially the artificial neural network, can approximate complex functions. However, they lack interpretability and struggle when data is scarce or in extrapolation regimes. Empirical models assume that the data is of sufficient granularity and quantity to define the system. Today's smart sensors and sophisticated technologies have paved the way for the availability of tremendous data for cyber-physical industrial systems. However, selecting legitimate data for training in process systems is still a daunting task. The exploratory data techniques are of great help in finding potential data for intended functions. Nonetheless, sensor breakdown, process shutdown, malfunctioning of equipment, random fluctuations, incorrect calibration, inconsistent sampling frequencies, and data entry errors due to human factors cause various quality-related issues such as outliers, missing information, imbalanced data, poorly labeled data, incorrectly mapped properties, noise, and inconsistencies. The model will fail at making a reliable assessment if the training data is only representative of a subsection of the processing system (Alvarez et al., 2009). Knowing the system boundaries of a process usually requires prior knowledge elicited from process industry practitioners.

Industrial processes are often described as *data-rich but information-poor*. The imbalanced and sparse data are also a growing concern for the training of data-driven models of oil and gas systems. These systems usually have a high volume of data of normal conditions while fewer representative samples of the anomalies. The industrial systems have a large quantity of process data from

conventional measurements (pressure, temperature, flow rate) but a lack of data on stream compositions, and other sophisticated measurements.

First principle models on the other hand rely on system understanding to compensate for the lack of data (Fisher et al., 2020). Mechanistic models are formalized using conservation laws, domain knowledge, physical principles, and phenomenological behaviors. Because of this, they have a greater potential for extrapolation compared to empirical models. Such models, however, need the correct specification of all interactions between variables and system parameters which is not feasible in the modern era's complex industrial systems.

Many efforts are made for bridging the gaps in the data-based and mechanistic models. However, there is still a lack of literature on integrating the process knowledge such as material and energy balances, physical laws, chemical kinetics, and expert knowledge in the training of data-based models. There is an ongoing debate on how much domain knowledge is necessary for efficient learning. At one extreme is "blank slate" learning with no domain knowledge to the entirely ridden by the physical laws at the other extreme.

The motivation of this work is to reconcile data-driven models with mechanistic modeling to develop hybrid models that can incorporate prior knowledge, and provide credible predictions. This work contributes to the following.

    i.   *Devising process dynamics-guided deep neural network formalism:* We have developed a process dynamics-guided deep neural network model for improving the performance of the system. This has been achieved by adding an additional layer for establishing the process dynamics such as material and energy balance equations and process heuristics.

ii. *Evaluation of the proposed dynamics-guided deep neural network (PDNN):* We have examined the proposed model on three virtual processes, simple linear, nonlinear, and moderately complex processes. The model has been examined for detecting low-quality distilled fractions of a binary separator. The proposed hybrid model yielded improved results on reduced samples sizes that endorse better generalization ability of the process dynamics-guided neural network model. It has also been examined on forecasting scenarios in a transient phenomenon. We have studied reaction kinetics following systems of ordinary differential equations to study the conversion of reactants and products in a batch reactor.

The subsequent sections of this paper are organized as follows. Section 4.1.2 provides a brief survey of data-driven and mechanistic models with a focus on the deep neural network developed for process systems. Section 4.2 presents mathematical models for integrating the process dynamics in deep neural network models. The model outcomes are presented in Section 4.3 followed by a conclusion (Section 4.4).

4.1.2 Related Literature

Numerous data-based methodologies such as principal component analysis (PCA) and partial least squares (PLS) (Kaspar & Harmon Ray, 1993*;* Kourti & MacGregor, 1995; Qin, 1998), independent component analysis (ICA) (Stefatos & Hamza, 2010; Zhang & Qin, 2007; Zhang & Zhang, 2010), vine copula-based methods (Cui & Li, 2020; Ren et al., 2017; Yu, Khan & Garaniya, 2015a), Gaussian mixture model (GMM), (Li, Qin & Yuan, 2016; Thissen et al., 2005; Yu, 2013; Yu & Qin, 2008, 2009) have been studied for fault detection and diagnostic methods. Many machine learning techniques such as artificial neural network (ANN) (Agatonovic-Kustrin & Beresford, 2000; Li et al., 2017; Park & Sandberg, 1991; Schmidhuber, 2015; Zhang et al., 2020; Zhao et al.,

2018), support vector machines (SVM) (Burges, 1998; Cortes & Vapnik, 1995; Widodo & Yang, 2007), case-based reasoning (CBR) (Zhao et al., 2017), and classification and regression tree (CART) (Lawrence & Wright, 2001) have been successfully tested for fault detection and diagnostic methods for process systems. Alauddin, Khan, Imtiaz, and Ahmed (2018) classified the evolution and development of research on data-driven FDD in four categories, namely, formulation of basic data-driven algorithms, advancement of the algorithms, applications of the algorithms in process systems, and development and application of advanced hybrid techniques in process systems. Some of the comprehensive reviews on data-driven FDD includes: (Alauddin et al., 2018; Arunthavanathan et al., 2021; Dai & Gao, 2013; Gao et al., 2015; Ge, 2017; Qin, 2012; Tidriri et al., 2016; Yin et al., 2012, 2014).

Mechanistic modeling has been a major part of Chemical Engineering and a central activity of Process Systems Engineering (PSE) for several decades. These models have a stronger ability for generalization and extrapolation. The synergy between the two approaches has been gaining attention, by either redesigning the model's architecture, augmenting training datasets with simulated data, or including physical principles as constraints in the cost function. The hybrid models are also referred to as a gray-box model (GBM) (Acuña et al., 2013), integrated neural network (Mavkov et al., 2020), semi-mechanistic (Lima & Saraiva, 2007), semi-parametric (McBride et al., 2020; von Stosch et al., 2014a), informed machine learning (von Rueden et al., 2021), domain adaptive (Liu et al., 2020; Schuld et al., 2014), theory-guided data science (Downton et al., 2020; Karpatne et al., 2017), and physics-guided neural networks (Hu et al., 2021; Murphy & Kerekes, 2021; Wang et al., 2020). Hybrid models combine the flexibility and scalability of machine learning while respecting the physics of the underlying systems. The hybrid models can result in improved prediction accuracy, better calibration properties, enhanced

extrapolation properties, and better interpretability (Chaffart & Ricardez-Sandoval, 2018; Psichogios & Ungar, 1992; Reichstein et al., 2019; Svendsen et al., 2021; Thompson & Kramer, 1994; Van Lith et al., 2003; von Stosch et al., 2014b).

The distinct hybrid formalism could be achieved via *Proxy* (one model acts as a surrogate to the other), *complement* (the solution is a combination of the two models), *supplement* (a model provides a correction for the other model), *embedment* (a model is embedded within the other model), *integrate* (output of a model serves as an input for the other model), and *inspiration* (the structure of a model is developed from the knowledge provided by the other model) (Fisher et al., 2020; von Stosch et al., 2014b). Thompson and Kramer (1994) divided techniques to combine prior knowledge with neural networks in design approaches (modular, serial, and parallel) and training approaches (objective function and constraints). The design approach uses prior knowledge as a basis for selecting a network.

Integrated mechanistic knowledge in the form of physics-based priors has been rendered in many data-based models including kernel machines (Acuña et al., 2013), Bayesian non-parametric approaches such as Gaussian Process (GP) regression (Camps-Valls et al., 2018; Raissi et al., 2017; Svendsen et al., 2018) and latent force models (LFMs) (Alvarez et al., 2009). Camps-Valls et al. (2018) presented an automatic Gaussian process emulator (AGAPE) that approximates the forward physical model using Bayesian optimization and look-up-table. Latent force models (LFMs) were devised to incorporate physical knowledge encoded in differential equations into a multi-output GP model. LFMs can transfer information across time series, handle missing observations, infer explicit latent functions forcing systems, and learn parameterizations. Numerous filter based estimators such as Kalman filter (Santos et al., 2021; Sohlberg, 2003), extended Kalman filter (EKF) (Destro et al., 2020), and unscented Kalman filter (UKF) (Kreuzinger

et al., 2008; Simutis & Lübbert, 2017) have been employed in model predictive control and state estimation of process systems. Several comprehensive reviews on hybrid methods applied to process systems have been presented (Bikmukhametov & Jäschke, 2020; Sansana et al., 2021; von Stosch et al., 2014a; Xiong & Jutan, 2002).

Table 4.1 presents a hybrid formalism for integrating process knowledge in a neural network framework. Psichogios and Ungar (1992) employed a hybrid model where neural network component processes parameters of the first principle model. Thompson and Kramer (1994) studied a hybrid model based on a parametric model (that compensates for sparse data) and a neural network that accounts for uncertainties and biases. Xiong and Jutan (2002) investigated a hybrid model-based control strategy using a parallel structure where a neural network was used to compensate model mismatch of detailed and approximate mechanistic models. Georgieva et al. (2003) employed a hybrid model in an industrial fed-batch evaporative crystallization process for predicting size distribution in a refining process. Mass and energy balances were captured in the mechanistic sub-model whereas the data-driven sub-model described growth rate, nucleation, and agglomeration parameters. Oliveira (2004) proposed a hybrid formalism with parallel and fused multiplication functions. Stewart and Ermon (2017) formulated a new approach to supervising neural networks by specifying mechanistic constraints in the output space. Azarpour et al. (2017) developed a generic framework based on the first principle model and an artificial neural network to study catalyst deactivation of fixed-bed catalytic reactors (FBCRs). Muralidhar et al. (2018) illustrated how incorporating domain constraints directly into the loss function can improve the learning of parameters. Hendriks et al. (2020) presented a neural network-based model that can explicitly satisfy known linear operator constraints.

Wu et al. (2020) proposed three distinct hybrid formalisms based on a recurrent neural network to integrate physical knowledge in data-driven models.

Many studies conferred that meta-learning can significantly improve the performance of deep neural networks. Van Lith et al. (2003) augmented a simple physical framework with fuzzy logic. Lee and Kang (2007) formulated a modified back-propagation neural network (BPN) based meta-model that ensures constraint feasibility of the approximate optimal solution. Alauddin et al. (2020) presented an ANN-based fault detection model that was tuned using a variable mosquito flying optimization technique for maximizing fault detection rate and minimizing false alarm rate.

Table 4.1: Hybrid formalisms for integrating process knowledge in neural network models of process systems

	Study	Hybrid Formalism	Case study
	(Psichogios & Ungar, 1992)	Process parameters of the first principle model were determined by neural network	Fed-batch bioreactor
	(Thompson & Kramer, 1994)	Parametric model that compensates for sparse data and neural network accounting for uncertainty and bias	Penicillin fermentation
	(Xiong & Jutan, 2002)	ANN was used to compensate for model mismatch of detailed and approximate mechanistic models.	Batch reactor; continuous stirred tank reactor
	(Bollas et al., 2003)	Neural model was used to refine the plant model prediction	Fluid catalytic cracking
First principle models	(Georgieva et al., 2003)	Mass and energy balances were captured in the mechanistic sub-model the data-driven sub-model described the growth rate, nucleation, and agglomeration parameters.	fed-batch evaporative crystallization process
	(Oliveira, 2004)	Parallel and fused mechanistic part and a data-driven part.	Recombinant protein and baker's yeast production
	(Safari et al., 2014)	Neural network fused the outputs of multiple Kalman filters	Industrial sensor fusion
	(Yang et al., 2020)	Integrating deep neural network with a physical lumped kinetic model	Automated FCC process.
	(Chen & Ierapetritou, 2020)	Serial, parallel, and combined structures of hybrid models.	simulated reactor mode

	(Wu et al., 2020)	Physics-based recurrent neural network (RNN)	model predictive control
	(Ghosh et al., 2021)	Data-driven model learns the residuals from the mechanistic model; mechanistic nonlinear model approximated by surrogate linear model and learning of residual using a data-driven model	Batch crystallization
	(Tan & Li, 2002)	Mechanistic sub-model based on momentum balance and the data-driven model using Padé approximation	Hydraulic nonlinear system
	(Van Lith et al., 2003)	Simple physical framework augmented with fuzzy logic. Information about the dynamic behavior is incorporated in the form of prior knowledge	Experimental batch distillation column,
Meta-modeling	(Lee & Kang, 2007)	Modified back-propagation neural network (BPN) based meta-model that ensures constraint feasibility of the approximate optimal solution.	Standard structural problems
	(Alauddin et al., 2020)	ANN-based fault detection model was tuned using a variable mosquito flying optimization (V-MFO) technique for maximizing fault detection rate (FDR) and minimizing false alarm rate (FAR).	Tennessee Eastman process

4.2 The proposed model

We have developed a process dynamics-guided deep neural network model for improving the generalization of the system. Fig. 4.1 illustrates the core idea behind mechanistic, data-driven, and integrated models. Suppose that we have to predict the activity of a person at a given time window using the mechanistic and data-based models. The mechanistic models will be based on a predefined architecture for distinct activities such as sleep patterns, eating habits, and other physiological behavior. There will be a designated structure of the work hours, recreation, and social engagements. The added responsibility of the family adds other constraints for predicting the activity. An effective mechanistic model will also delineate the effect of seasonality and aging

on the activity of the person. The data captured by the tracking system will be allocated to any one of those predefined modules. The mechanistic model requires only a few data to completely specify the parameters of the model.

This task can also be accomplished by a robot that is unaware of any notion of human physiology, work environment, and social culture. The robot can predict the activity of the person in the stated time window with a complex virtual model based on history. The robot's black box model will require myriads of relevant data for training to accomplish this task.

The integrated model (Fig. 4.1c) can harness the benefits of both mechanistic and data-driven models. It can lead to an improved model with better generalization (i.e., better results on unseen data) without specifying all the interactions of the complex systems. The integrated model can also aid innovation and new dimensions of explorations. With this brief background, the next section presents the proposed process dynamics guided deep neural network.

4. 2.1. Process dynamics-guided neural network (PDNN)

The artificial neural network has been inspired by biological neural systems. A standard ANN consists of many processors, also called neurons that generate a sequence of real-valued activations (Agatonovic-Kustrin & Beresford, 2000; Schmidhuber, 2015). ANN can assimilate highly complex relationships between several variables by learning the nature of the dependency between input and output variables. The learning process is based on pattern recognition characteristics of a neural network.

A processing system can be characterized by three types of knowledge about the processes: mechanistic, heuristic, and knowledge within process data. They could be represented by distinct mathematical models ranging from simple algebraic equations to complex equations such as

systems of ordinary differential equations (ODEs), differential-algebraic equations (DAEs), and partial differential equations (PDEs). We have categorized this information into four levels as illustrated in Fig. 4.2. The first level is related to conservation laws such as material balance, energy balances, and universal laws. This is strictly followed by any system and subsystem of industrial processes. Level 2 captures regulations and bound constraints. For example, regulations on emission, the upper limits on the tolerable region in the ALARP-based design, bounds on physical parameters such as non-negative density. Level 3 captures the details of process dynamics and chemical processing operations such as reaction, separation, heat exchange, and waste treatment. The various chemical processing operations from the upstream (i.e., exploration, drilling, production) and downstream (such as mixing, separation, heat exchange, refining, waste minimization) are also taken into account in this layer. This layer can assimilate acquired knowledge from fluid mechanics, heat transfer, mass transfer, reaction engineering, process control, process safety, plant design, and economics. The layer aims to incorporate complex equations such as partial differential equations, fundamental correlations, and nonlinear equations.

Level 4 is related to applying heuristics and ergonomics. This addresses the mismatch between people, tools, and the working environments. This layer can aid some quick and relatively inexpensive designs heuristics in devising data-based models of processing facilities.

We are defining process relations and stoichiometry by adding an additional layer in the artificial neural network structure (Fig. 4.3). The proposed model is a multi-layer network comprising a process info layer, an input layer, hidden layer(s), and an output layer. The model is different from the standard ANN in terms of defining correlations and dependability relation in the first layer of the neural network model of a process system. We can incorporate some or all mechanistic

equations of process systems in the process dynamics layer. The model parameters are iteratively determined using the forward prediction and backward propagation through the network layers. The parameter learning of the network could broadly be summarized under three steps:

Step I: Forward propagation: The forward propagation of the network facilitates predictions based on the input and the network parameters. Eqs 4.1-4.5 represent the forward-pass model, where X is the input variables, W_i, W_1 and W_2 denote the weights and a_1 is the activation. The activation function a_1 can be a linear function, hyperbolic tangent, logistic function, or a rectified linear unit (ReLU). The model output 'y' (Eq. 4.6) could be a linear, sigmoid, or a softmax function depending on the nature of the problem i.e., regression or classification.

$$z_i = f(X.W_i) \qquad (4.1)$$
$$z_1 = f(z_i.W_1) \qquad (4.2)$$
$$a_1 = f(z_1) \qquad (4.3)$$
$$z_2 = a_1.W_2 \qquad (4.4)$$
$$y = f(z_2) \qquad (4.5)$$

Step II: Calculation of cost function: The loss function evaluates "goodness" of the model predictions. It is a measure of the cost caused by incorrect predictions. Eq. 4.6 presents the loss function, it can be different for regression and the classification tasks.

$$L = f(t_i, y) \qquad (4.6)$$

Step 3: Backpropagation and weight update: The network parameter Θ is determined by using the backpropagation based on the input features x, label t, and data quality. The backpropagation optimizes the network parameters for arbitrarily mapping inputs to outputs. The backpropagation is based on the chain rule to compute the derivatives of the loss function with respect to the network parameters (Aggarwal and Murty, 2021; Hinton, 1986).

$$\frac{\partial L}{\partial W_i} = \frac{\partial L}{\partial a_2} \frac{\partial a_2}{\partial z_2} \frac{\partial z_2}{\partial a_1} \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial W_1} \frac{\partial z_i}{\partial W_i} \qquad (4.7)$$

$$\frac{\partial L}{\partial W_1} = \frac{\partial L}{\partial a_2} \frac{\partial a_2}{\partial z_2} \frac{\partial z_2}{\partial a_1} \frac{\partial a_1}{\partial z_1} \frac{\partial z_1}{\partial W_1} \qquad (4.8)$$

$$\frac{\partial L}{\partial W_2} = \frac{\partial L}{\partial a_2} \frac{\partial a_2}{\partial z_2} \frac{\partial z_2}{\partial W_2} \qquad (4.9)$$

Eq. 4.10-4.12 presents the *weight update equation* where η is the *'learning rate'* that controls the influence of gradient on the weight update. To adequately explore the search space, we have employed a step decay adaptive learning rate. Eventually, the weight of the process dynamics layer can be re-adjusted based on the process knowledge. The re-adjusted weight is fed to the forward and the backward loop. This cycle continues until the convergence of the desired accuracy is achieved.

$$W_1 = W_1 - \eta \frac{dL}{dW_1} \qquad (4.10)$$

$$W_2 = W_2 - \eta \frac{dL}{dW_2} \qquad (4.11)$$

$$W_i = W_i - \eta \frac{dL}{dW_i} \qquad (4.12)$$

The process dynamics and other inputs are incorporated as follows:

A. *Incorporating material balance equations without reactions:* The material balance equations of process design is usually algebraic equations. The simple linear equations can be $\sum m_i = 0$. Suppose we have an equation $x_1 = x_2 + x_3$. This is induced as follows-

- o Initialize parameters
- o Calculate network output using the forward propagation
- o Calculate the loss function
- o Revise weights using backpropagation
- o Tune weights of the process dynamics layer to follow the constraint(s)
- o Repeat these steps until the desired convergence is achieved

B. *Handling differential equations*: Process dynamics is usually described using differential equations. For example, a material balance equation with a chemical reaction results in an ordinary differential equation. The artificial neural network can handle the dynamic modeling based on a sequence of transformations to hidden states (Eq. 4.13). The discrete layers of a neural network can be represented using continuous dynamics leading to a differential equation (Eq. 4.14). Thus, a differential equation can be employed as a layer of a neural network (Chen et al., 2018; Rackauckas et al., 2020).

$$h_{t+1} = h_t + f(h_t, \theta_t) \tag{4.13}$$

$$\frac{dh_t}{dt} = f(h_t, t, \theta_t) \tag{4.14}$$

The loss function at any point t1 can be represented by Eq. 4.15.

$$L(z(t_1)) = \int_{t_0}^{t_1} f(z(t), t, \theta) dt \tag{4.15}$$

The parameters are iteratively determined by solving the ODE during the forward propagation and adjoint ODE during the backpropagation. Eq. 4.16-18 present the differential with respect to parameters, where, $a(t) = -\frac{\partial L}{\partial z(t)}$ (Surtsukov, 2021).

$$\frac{dL}{dz(t_0)} = \int_{t_1}^{t_0} a(t) \frac{\partial f(z(t), t, \theta)}{\partial \theta} dt \tag{4.16}$$

$$\frac{dL}{d\theta} = \int_{t_1}^{t_0} a(t)^T \frac{\partial f(z(t), t, \theta)}{\partial \theta} dt \tag{4.17}$$

$$\frac{dL}{dt_0} = \int_{t_1}^{t_0} a(t)^T \frac{\partial f(z(t), t, \theta)}{\partial \theta} dt \tag{4.18}$$

C. *Handling miscellaneous constraints*: The knowledge from operators, regulatory authorities, and other added constraints can be addressed by adding them as penalty functions in the loss function (Eq. 4.19). The objective of training a network is to find the parameters that minimize the loss function. The penalty function penalizes constraints violations i.e., the effective loss would be higher if the constraints are violated. The iterative process of forward and backward propagations will ultimately result in the optimal set of parameters while satisfying the constraints within the desired tolerance.

$$L_{Eff} = L\big(z(t)\big) + P * g(z_t) \qquad\qquad (4.19)$$

Tracking system

(A)
Mechanistic Model

(Few data for determining the parameters of the physical model)
- Better generalization
- Valid outcomes for extrapolations
- Easy interpretability

Aging Effect

Social life

Recreation

Working hours

Physiology

(B)
Data-driven Model

(Bulk data to determine parameters of data-driven models)
- No week assumption
- No need to specify all interactions beforehand
- Scope for innovation

(Few data for determining the parameters of the physical model)
- Better generalization
- Valid outcomes for extrapolations
- Easy interpretability

Aging Effect

Social life

Recreation

Working hours

Physiology

(C)
Integrated Model

(Many data for determining the parameters of the data-driven models satisfying the process dynamics)
- No week assumption
- No need to specify all interactions beforehand
- Scope for innovation

Fig. 4.1: Mechanistic-based data-driven model
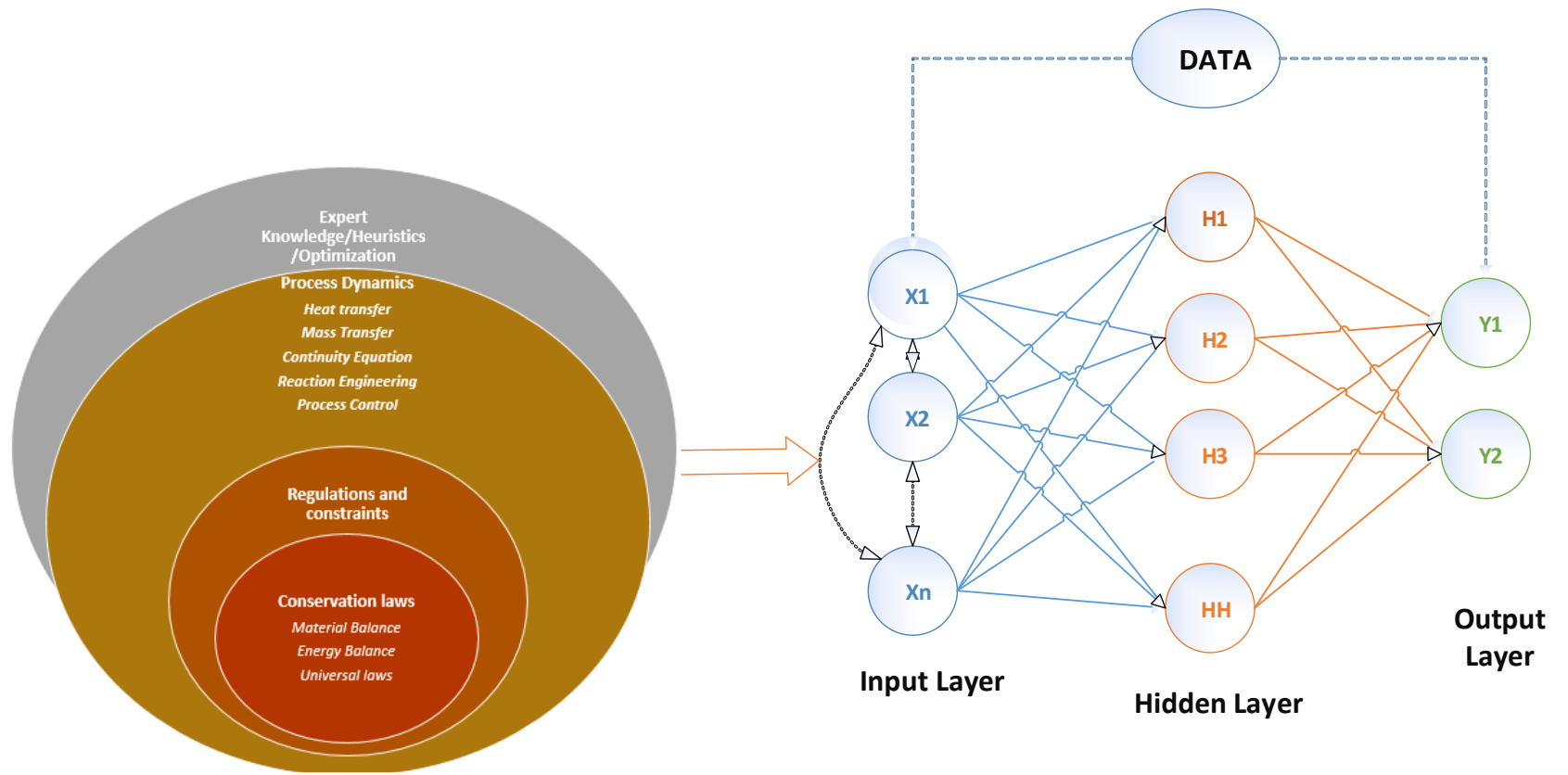
Fig. 4.2: Process dynamics-guided neural network model of a process system
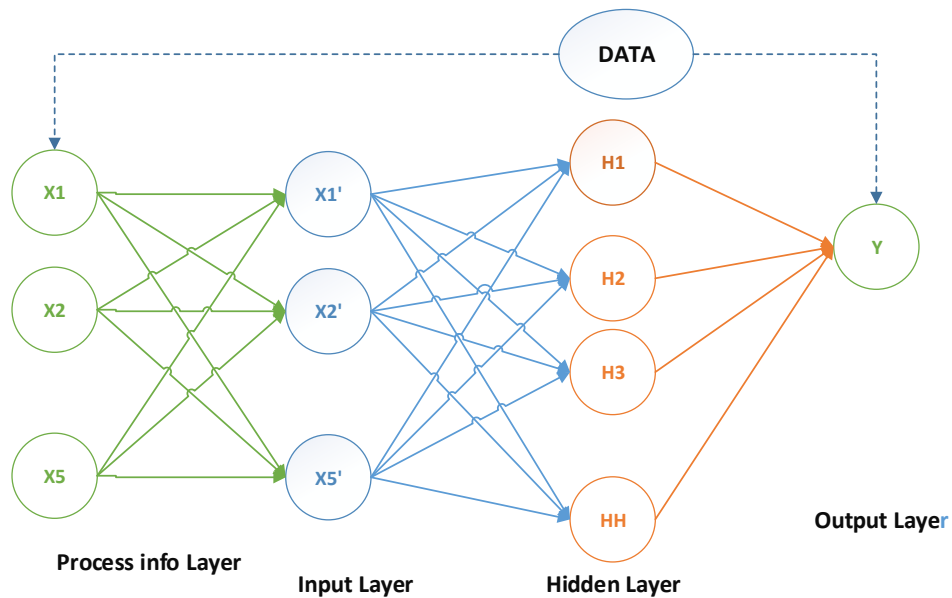
Fig. 4.3: Defining process dynamics and expert knowledge in the first layer of a neural network model of a process system

4.3 Results and Discussion

The proposed model has been evaluated in regression and classification tasks which are widely used in design, control, optimization, scheduling, and risk analysis of process systems. Regression is used to deduce the relationship between interacting variables and process parameters. The chemical engineering applications of regression include estimating reaction kinetics, equilibrium constant, adsorption isotherms, catalyst design, loading rate, drying time, heat transfer coefficients, product design, and predicting fluid properties and streams composition. The classification is exploited for anomaly detection, fault diagnosis, and decision making in distinct scenarios of risk associated with a process system. The process dynamics-guided neural network (PDNN) model has been studied on steady states and transient operations. Sections 4.3.1 and 4.3.2 present the evaluation of the model on regression and classification operations assuming a steady-state whereas Section 4.3.3 analyzes the transient behavior of a batch reactor.

111

4.3.1 Evaluation of process dynamics-guided neural network (PDNN) on regression tasks

Fig. 4.4 presents a simplified block diagram of a simple chemical processing system. It comprises two splitters and a complex processing system with multiple operations. The process output 'y' is a function of measured variables $x_1, x_2, x_3, x_4,$ and x_5 (*Eq.* 4.20). We have simulated three virtual processes, a simple linear process, a nonlinear process, and a moderately complex process (Eq. 4.20-23). The function f_1 is a simple linear process, f_2 is a nonlinear process of second-order while f_3 is a complex nonlinear process capturing the orientation along with the variable interactions. C1 and C2 present constraints inferred from the material balance of the system.

$$y = f(x_1, x_2, x_3, x_4, x_5) \qquad (4.20)$$

$$f_1 = x_1 - x_2 - x_4 \qquad (4.21)$$

$$f_2 = (x_1 + x_2) * (1 + x_3) + x_3 * x_4 \qquad (4.22)$$

$$f_3 = 10x_1 \sin(x_2) + x_3\{1 - \exp(x_4)\} \qquad (4.23)$$

$$C1: \quad x_1 = x_2 + x_3 \qquad (4.24)$$

$$C2: \quad x_1 = x_2 + x_3 \qquad (4.25)$$
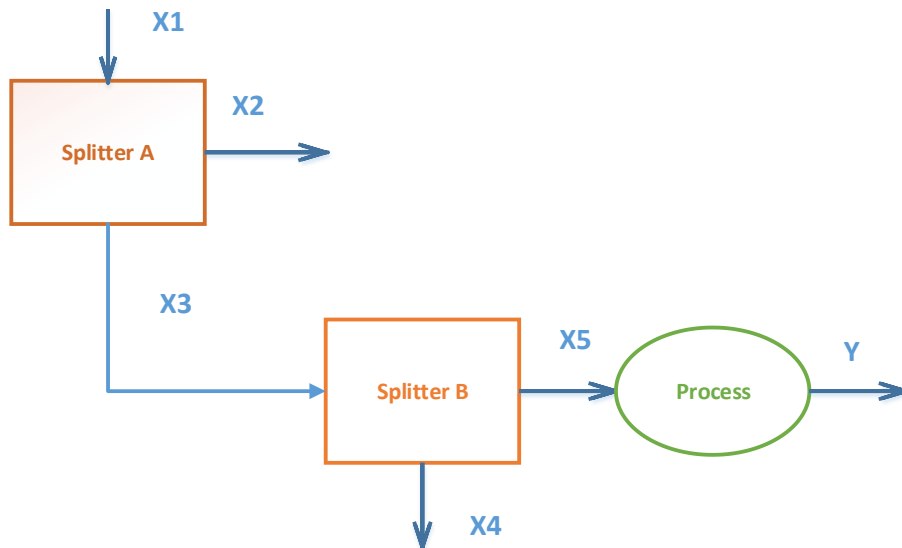
Fig. 4.4: Block diagram of a simple chemical processing system

The test models have been evaluated on noisy data generated from the virtual functions f_1, f_2, and f_3 in terms of the root mean squared error (RMSE) which is a measure of the difference between the calculated and the observed values ($RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{estimated,i} - y_{true\ value,i})^2}$). We divided data in to training and test samples in a ratio of 70: 30. Fig. 4.5 presents the accuracy of the test models on training samples (Fig. 4.5A) and on the test data (Fig. 4.5B). The **ANN** represents the standard ANN model based on data without considering the balance equations. The **PDNN-1** represents the proposed process dynamics-guided neural network model incorporating constraint **C1**, while **PDNN-2** denotes the proposed process dynamics-guided neural network model for incorporating both constraints **C1** and **C2**.

The RMSE values of the standard ANN model and the proposed process dynamics-guided neural network models (PDNN-1, and PDNN-2) for the nonlinear process f_2 were found 0.1247, 0.1234, and 0.1233 respectively on the unseen data. We did not find any significant difference in the accuracy of the models on the training samples. However, process dynamics-guided neural network models yielded improved results on the test data. This speaks for the better generalization ability of the proposed process dynamics-guided neural network models. Table 4.2 presents the accuracy of test models of three processes of distinct complexities on training samples and unseen data. Here again, the process dynamics guided deep neural network is producing improved outcomes, especially on the test samples. Hence, the proposed PDNN is effective in handling regression tasks of process systems. The next section presents the model performance on a classification task of finding distillates of the desired composition in a binary distillation.

Fig. 4.6 displays the history of parameter learning of test models. We can see that the loss function in terms of the root mean squared error (on the y-axis) is improving with increasing epochs. The

process dynamics guide in effective learning of the parameters. We can also notice that every additional feeding from the knowledge of the process dynamics improves the model. Thus, **PDNN-2** that incorporates both constraints *C1* and *C2* perform better than the **PDNN-1** that accounts for constraint *C1* only. However, the difference is negligible at increasing learning rates as shown in Fig. 4.6B-C. It can also be noticed that the additional constraints slow down the convergence especially at lower learning rates. This can be compensated by increasing the number of epochs.
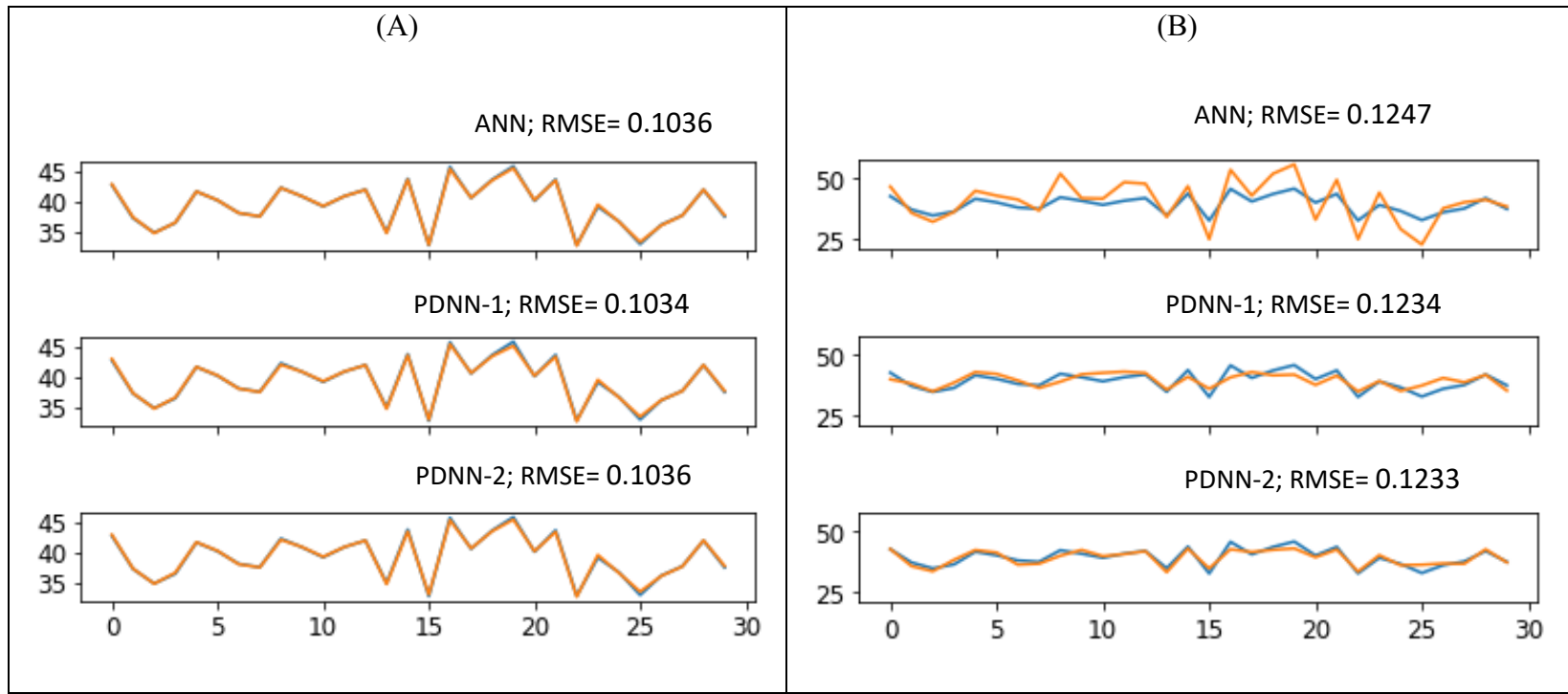
Fig. 4.5: The accuracy of the test models on training and test data (A) training data, (B) test data

Table 4.2: Accuracy of test models on training and test data of virtual processes

Test functions	RMSE on training data			RMSE on test data		
	ANN	PDNN-1	PDNN-2	ANN	PDNN-1	PDNN-2
$f_1 = x_1 - x_2 - x_4$	0.0016	0.0017	0.0017	0.0016	0.0014	0.0014
$f_2 = (x_1 + x_2) * (1 + x_3) + x_3 * x_4$	0.1036	0.1044	0.1036	0.1247	0.1234	0.1233
$f_3 = 10x_1 \sin(x_2) + x_3\{1 - \exp(x_4)\}$	2.2771	2.2771	2.2772	2.0969	2.0635	2.0417
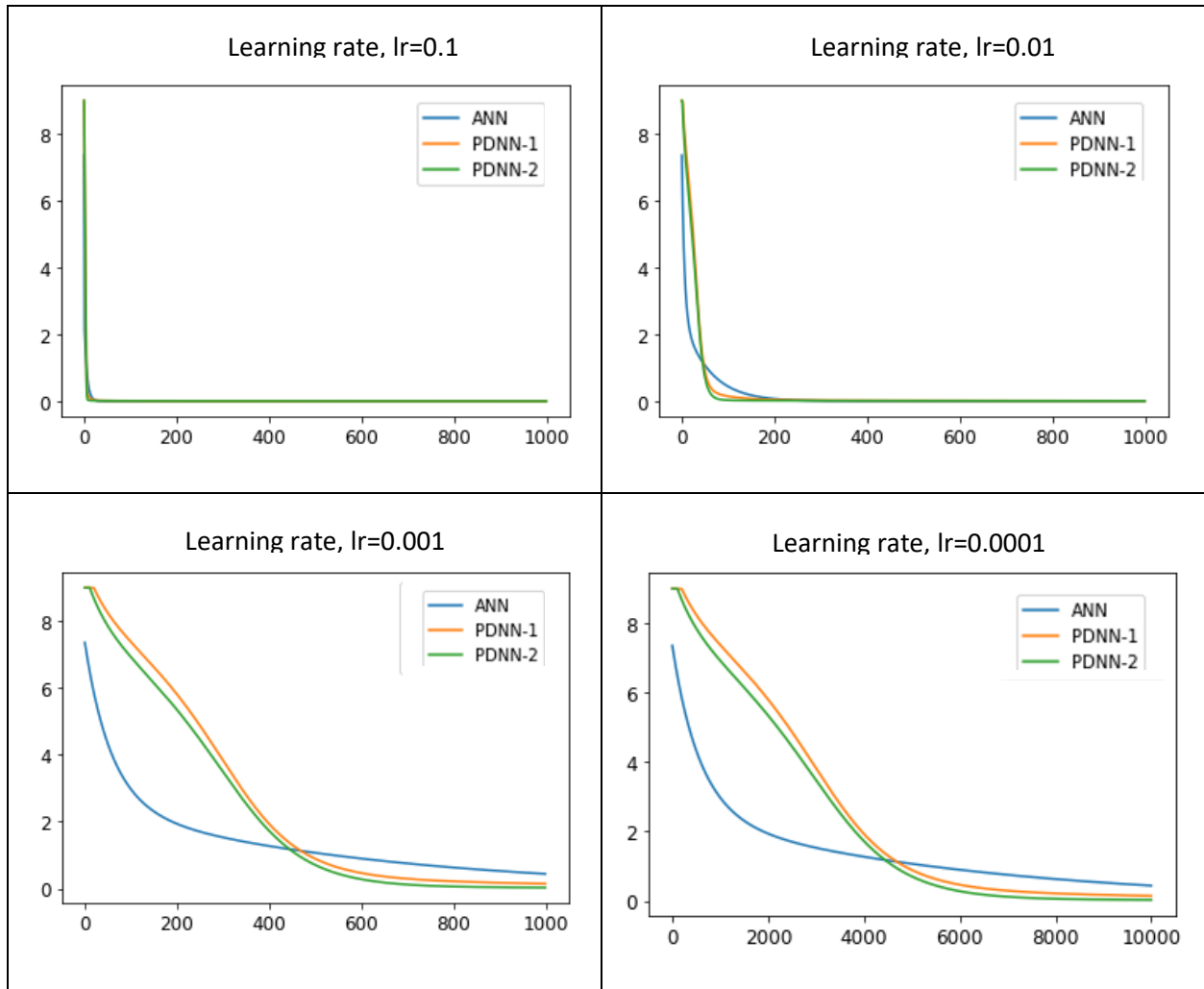
Fig. 4.6: Convergence of the calculated value with the target value with epochs at different learning rates

4.3.2 Evaluation of the process dynamics-guided neural network (PDNN) on a classification task

Distillation is a widely used unit operation of petroleum refining, chemical processes, and allied industries. Fig. 4.7 presents a binary distillation column comprising a vertical shell, a condenser, and a reboiler. The feed tray divides the column into the top (enriching) and bottom (stripping) sections. The feed flows down the column where it is collected at the bottom. The vapor rises to the top of the column, cooled by a condenser and removed as a distillate. A fraction of this liquid, also known as reflux, is recycled back to the column through enriching section. The quality of

distillates is a function of relative volatilities of components, their activity coefficients, the ratio of liquid-phase to the vapor-phase flow rates, and the ratio of surface area to the volume of liquid in the distillation column.

The liquid and vapor leaving the tray are in equilibrium with the vapor and liquid entering the tray. The moles of vapor condensed equates the moles of liquid vaporized assuming constant molar flow rates of the vapor and liquid in each section of the column and negligible heat losses to and from the column. Eq. 4.26 and Eq. 4.27 respectively represent the mass balance and the component balance in the column. Eq. 4.28-4.29 represent material balances at sections above the feed; Eq. 4.30-4.31 represent the balances below the feed, and Eq. 4.32-4.33 present the mass balance at feed location. The notations are as follows: V= vapor flow rate; L= liquid flow rate; D= distillate flow rate; B=bottoms flow rate; F= feed flow rate; X_F, X_D, X_B = fraction in feed, distillate, and bottoms, q= quality of feed, and R= reflux ratio.

$$F = D + B \tag{4.26}$$

$$FX_F = DX_D + BX_B \tag{4.27}$$

$$V_T = L_T + D \tag{4.28}$$

$$y_{n+1}V_{n+1} = x_n L_T + X_D D \tag{4.29}$$

$$L_B = V_B + B \tag{4.30}$$

$$x_m L_B = y_{m+1}V_B + B \tag{4.31}$$

$$V_T = V_B + (1 - q)F \tag{4.32}$$

$$L_B = L_T + qF \tag{4.33}$$

$$D = \frac{V}{R + 1} \tag{4.34}$$
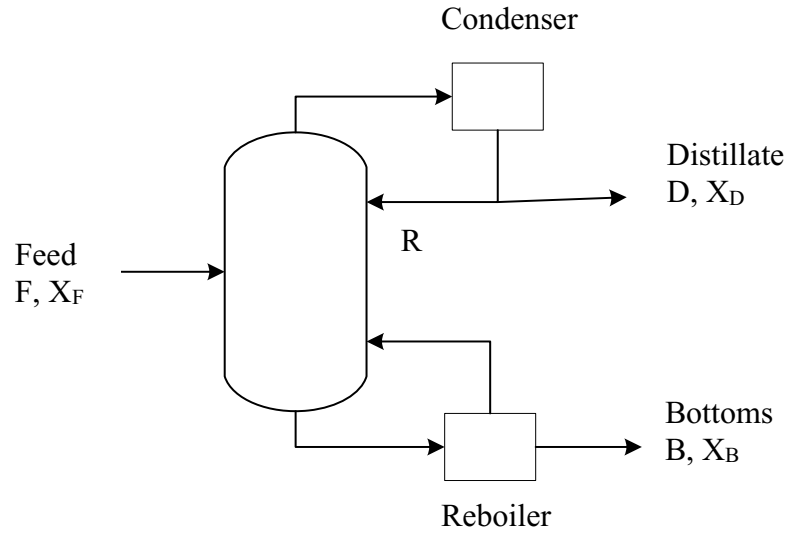
Fig. 4.7: A simplified representation of a binary distillation column

The proposed process dynamics-guided neural network model has been examined for detecting low-quality distillate in a binary separator. We have evaluated the performance of the model using a receiver operating characteristic (ROC) curve. The ROC curve helps in visualizing the trade-off between the false positive and the true negative rates of a classifier. Fig. 4.8 presents the receiver operating characteristic curves of detecting low-quality samples based on standard ANN and the proposed process dynamics-guided neural network (PDNN) models. Fig. 4.8A presents the ROC curves for detecting undesired samples on test data based on 1000 data (700 training data; 300 test data). The dotted diagonal in the ROC curve represents a random classification model that does not have any ability to distinguish between the two classes at any threshold. An effective detection system can distinguish desired and undesired samples well if its ROC curve will be farther away from the diagonal. The area under the ROC curve (AUC) represents the degree of separability. The higher the AUC, the better the model is at predicting the distinct classes. The AUC is classification-threshold-invariant, it measures the quality of the model's predictions regardless of the

classification threshold. We can see that both of the test models record higher accuracy of 99 % in detecting the desired and the undesired samples.

We also assessed the performance of test models on a reduced number of training samples. The performance of both models was deteriorated by reducing the test training data. However, the proposed process dynamics-guided neural network model maintained higher accuracy at reduced samples. The accuracies of the PDNN and the standard ANN models (in parentheses) were recorded as 0.97 (0.95), 0.94 (0.92), and 0.92 (0.88) on corresponding reduced sample sizes to 200, 100, and 50. The comparative improved results of the PDNN further endorse the better generalization ability of the proposed model.
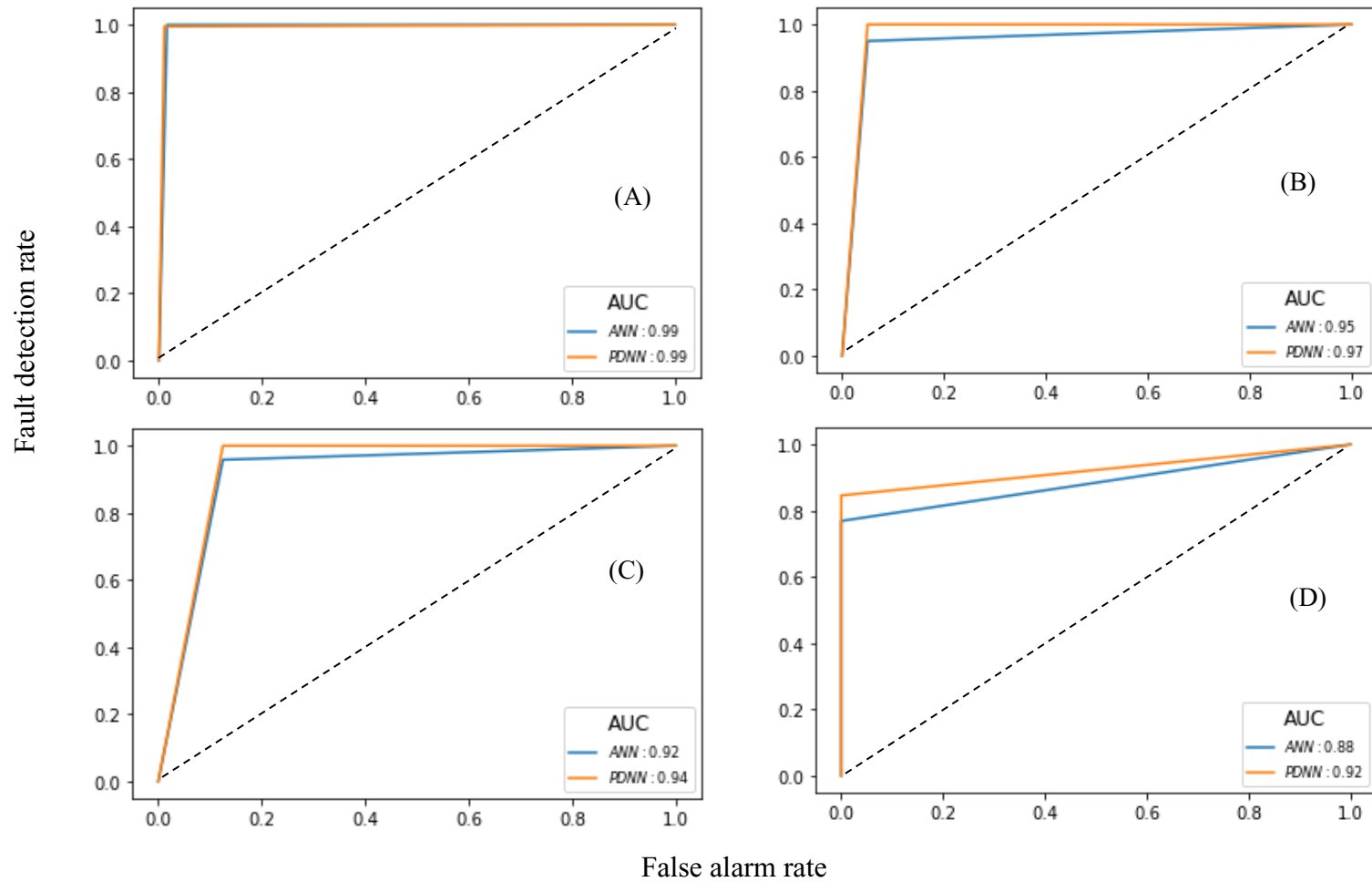
Fig. 4.8: The comparative ROC curves of the PDNN and standard ANN on for detecting undesired samples in distillates: (A) for 1000 samples (700 training, 300 testing) (B) for 200 samples (140 training, 60 testing) (C) for 100 samples (70 training, 30 testing) (D) for 50 samples (35 training, 15 testing)

4.3.3 Kinetic studies of a batch reactor using process dynamics guided neural network

Fig. 4.9 presents a schematic of a batch reactor with simultaneous reactions. The state variables involved in this process include concentration of reactant 'A', and the desired product 'B'. The reaction $A \xrightarrow{k_1} 2B$ denotes the conversion of the reactant 'A' to 'B' which is converted to 'C' via $B \xrightarrow{k_2} C$ mechanism. The conversion follows first-order kinetics with respect to species A and B. The kinetic is presented in Eq 4.36-36.

$$\frac{dA}{dt} = -k_1 A \qquad (4.35)$$

$$\frac{dB}{dt} = 2k_1 A - k_2 B \qquad (4.36)$$

$$A \xrightarrow{k_1} 2B$$
$$B \xrightarrow{k_2} C$$
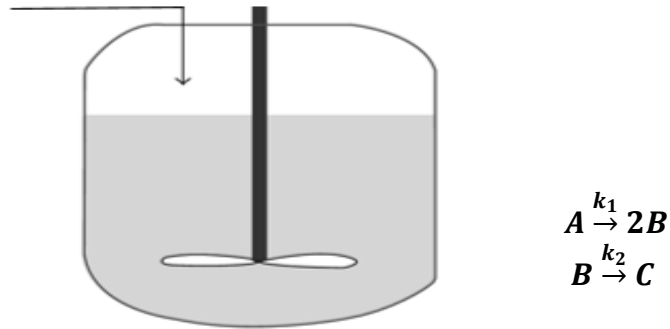
Fig. 4.9: A simplified representation of a batch reactor

The initial values of the state variables are as follows: $A_i = 10 \frac{mole}{m^3}$ and $B_i = 0 \frac{mole}{m^3}$. The kinetics of the reacting systems k_1 and k_2 are $1 \frac{mole}{m^3 \, day}$ and $0.1 \frac{mole}{m^3 \, day}$ respectively. Fig. 4.10 presents concentration profiles of A (Fig. 4.10A) and B (Fig. 4.10B) in the batch reactor using process dynamics-guided neural network (solid line) and standard data-driven neural network (dott-dashed line). The ordinates represent concentrations ($\frac{mole}{m^3}$) whereas the abscissa denotes time (in days). Both networks have been trained with data up to 10 days for predicting the profiles for the period of 15 to 25 days. The models are fitting well on training data; however, the proposed process dynamics-guided neural network model is more accurately forecasting scenarios of the

extrapolated data. We have also studied the predicting capability of test models on distinct labels of extrapolated samples. Thus, we analyzed three training windows (upto 5 days, 10 days, and 15 days) for predicting concentration profiles of 'A', and 'B' in the desired period of 15 to 25 days. The root-mean-squared errors in forecasting concentrations of the species using PDNN and the standard ANN models have been presented in Table 4.3. The accuracy of the models is deteriorating on predicting extrapolated samples; however, the PDNN yields better outcomes compared to the standard ANN. This states the effectiveness of the proposed process dynamics-guided neural network model in forecasting species concentrations of simultaneous reacting systems in a batch reactor.
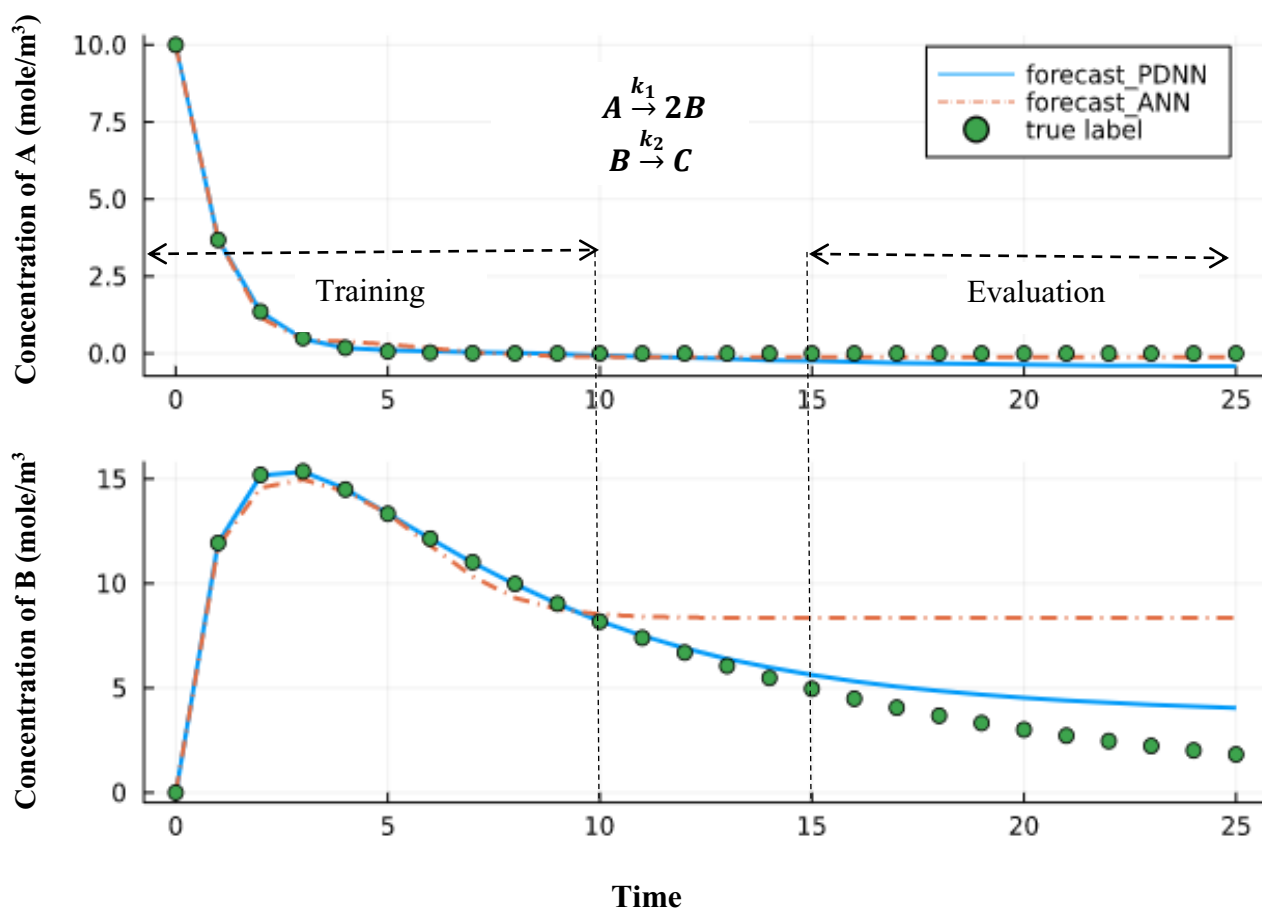
Fig. 4.10: Prediction of concentration of A and B in a batch reactor using process dynamics-guided neural network and standard data-driven neural network models

Table 4.3: Accuracy of test models for predicting species concentration in a batch reactor on extrapolated samples (i.e t=15 to t=25 days)

	RMSE		
	Training upto t= 5 days	*Training upto t= 10 days*	*Training upto t= 15 days*
PDNN	3.19	1.46	0.59
ANN	9.70	4.97	2.53

The study employed 10 neurons in the hidden layers of the test neural network models and Adam optimizer with an initial learning rate of 0.05. The Adam optimizer is based on adaptive estimates of lower-order moments that enables leveraging the power of adaptive learning to find individual learning rates for each parameter. The method is also appropriate for non-stationary objectives and noisy and/or sparse gradients. The findings can further be refined using hybrid optimization such as Adam followed by Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithms.

This study demonstrated the comparative effectiveness of the proposed process dynamics-guided neural network model in forecasting species concentrations in a batch reactor following systems of ordinary differential equations of the first order, the study can be extended to incorporate higher-order systems. Numerous process dynamics for example non-steady state diffusion in columns, axial dispersion in a tubular reactor, and non-homogeneous systems are represented by systems of partial differential equations. The process dynamics-guided neural network model can be expanded to study these systems. This work can further be extended to include alarms and operator's feedback to incorporate real-time data.

4.3 Conclusions

This work presented a process dynamics-guided deep neural network (PDNN) model to enhance model generalization by rendering process dynamics and field expertise. This has been realized by adding an extra layer for establishing the process dynamics such as material and energy balance equations. This formalism can results in improved generalization ability while facilitating scope for innovation for handling disturbances and process upsets. The proposed model has been evaluated on regression and classification tasks exhibiting steady-state behavior. The proposed PDNN model has also been employed to study the transient behavior of a batch reactor following systems of ordinary differential equations. The proposed hybrid model exhibited improved generalization ability on the unseen data and on the reduced samples for training. The process dynamics guided neural network model resulted in up to 10% improvement in predicting dependencies of virtual processes in mixing operations and up to 5% improvement in detecting low-quality distillates in a binary separator. The proposed model scores significant gains in forecasting the transient behavior of simultaneous reacting systems in a batch reactor as well.

Data-driven modeling, represented by machine learning (ML) and artificial intelligence (AI), is at the top of the Gartner hype cycle. Many experts believe that digital transformation is no longer simply a future option for processing systems, it's an urgent necessity. Mechanistic modeling has been a major part of Chemical Engineering and a key design tool of process systems engineering (PSE) since the publication of the celebrated Bird, Stewart, and Lightfoot's book, *Transport Phenomena* (Bird et al., 1960). We have several decades of expertise in designing processing facilities based on detailed knowledge about the system, including chemical kinetics, and transport (heat, mass, momentum) phenomena, materials, composition, boundary conditions, and thermodynamics properties. Several packages such as ASPEN Plus

(https://www.aspentech.com/en/products/engineering/aspen-plus), PRO/II (https://www.aveva.com/en/products/pro-ii-simulation/), ProSim (https://www.prosim.net/en/) , CADsim (https://www.aurelsystems.com/cadsim-plus/) , Chemcad (https://chemcad.co.uk/) , ChemPro (https://epcon.com/chempro.html), ChromWorks (https://www.ypsofacto.com/services-chemical-software-chromworks) , and DWSim (https://dwsim.fossee.in/) have been used in process design, monitoring and control. The incorporation of physics and domain knowledge can help to improve the robustness of data-driven models of complex industrial systems.

List of symbols and abbreviations

Symbols	Meanings
A	species A, concentration of A
$a(t)$	adjoint at time t
a_1	activation for the hidden layer
a_{1h}	h^{th} element of the hidden activation layer
a_2	activation for the output layer
a_{2j}	jth element of the output activation layer
A_i	initial concentration of A
ANN	artificial neural network
AUC	area under the receiver operating characteristic curve
B	bottoms flow rate
B	species B, concentration of B
BFGS	Broyden–Fletcher–Goldfarb–Shanno
Bi	initial concentration of B
BPN	back-propagation neural network
$C1, C2$	Constraints
CART	classification and regression tree
CBR	case-based reasoning
CI	classification index

D	distillate flow rate
DAEs	differential-algebraic equations
EKF	extended Kalman filter
F	feed flow rate
f_1, f_2, f_3	virtual functions
FAR	false alarm rate
FBCR	fixed-bed catalytic reactors
FDD	fault detection and diagnosis
FDR	fault detection rate
FL	fuzzy Logic
FN	true negative
FP	true positive
$g(z(t))$	constraint relations
GBM	gray-box model
GMM	Gaussian mixture model
GP	Gaussian process
HH	H^{th} neuron in hidden layer
$hi(t)$	hidden unit at time t
ICA	independent component analysis
IFCVAE	information concentrated variational auto-encoder
k_1	rate constant for conversion of A to B
k_2	rate constant for conversion of B to C
L	loss
L	liquid flow rate
L1	label 1
L2	label 2
LFMs	latent force models
Lr	learning rate
Number of samples	number of samples
ODEs	ordinary differential equations
P	penalty function
PCA	principle component analysis
PDEs	partial differential equations
PDNN	process dynamics guided neural network
PDNN-1	process dynamics-based neural network model incorporating constraint **C1**
PDNN-2	process dynamics-based neural network model incorporating constraint **C2**
PLS	partial least square
PSE	process systems engineering
Q	quality of feed
R	reflux ratio
ROC	receiver Operating Characteristic

S	the covariance matrix of process variables
sPCA	spherical principal component analysis
SVM	support vector machine
t_i	target value of the ith sample
TN	true negative
TP	true positive
UKF	unscented Kalman filter
V	vapor flow rate
W_1	weight matrix fro the hidden layer
W_2	weight matrix fro the output layer
W_{2ji}	weight of the jth neuron to i^{th} output
w_{FAR}	importance parameters for the false alarm rate
w_{FDR}	importance parameters for the fault detection rate
W_i	weight matrix for the input layer
X	training dataset
X_F, X_D, X_B	fractions of desired component in feed, distillate, and bottoms
X_{trust}	trusted centre of reliable dataset
$y_{estimated,i}$	estimated value for the ith sample
$y_{true,i}$	true label of the ith sample
z_1	hidden layer vector
z_2	output layer vector
x_k	k^{th} column of the input matrix
φ_d	standard deviation of d samples
η_d	tuning factor to adjust the calculation sensitivity
m_i	mass of ith species
θ_t	network parameters at time t
η	learning rate
Θ	network parameter

References

Acuña, G., González, J., Curilem, M., & Cubillos, F. (2013). A SVM gray-box model for a solid substrate fermentation process. *Chemical Engineering Transactions*, 35, 961–966. https://doi.org/10.3303/CET1335160

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN)

modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Aggarwal, M., & Murty, M. N. (2021). *Deep Learning. SpringerBriefs in Applied Sciences and Technology*. https://doi.org/10.1007/978-981-33-4022-0_3

Alauddin, M., Khan, F., Imtiaz, S., & Ahmed, S. (2018). A Bibliometric Review and Analysis of Data-Driven Fault Detection and Diagnosis Methods for Process Systems. *Industrial and Engineering Chemistry Research*, *57*(32), 10719–10735. review-article. https://doi.org/10.1021/acs.iecr.8b00936

Alauddin, M., Khan, F., Imtiaz, S., & Ahmed, S. (2020). A variable mosquito flying optimization-based hybrid artificial neural network model for the alarm tuning of process fault detection systems. *Process Safety Progress*, *39*(S1). https://doi.org/10.1002/prs.12122

Alvarez, M., Luengo, D., & Lawrence, N. D. (2009). Latent force models. In *Journal of Machine Learning Research* (Vol. 5, pp. 9–16).

Arunthavanathan, R., Khan, F., Ahmed, S., & Imtiaz, S. (2021). An analysis of process fault diagnosis methods from safety perspectives. *Computers and Chemical Engineering*, *145*, 107197. https://doi.org/10.1016/j.compchemeng.2020.107197

Azarpour, A., N.G. Borhani, T., R. Wan Alwi, S., A. Manan, Z., & I. Abdul Mutalib, M. (2017). A generic hybrid model development for process analysis of industrial fixed-bed catalytic reactors. *Chemical Engineering Research and Design*, *117*, 149–167. https://doi.org/10.1016/j.cherd.2016.10.024

Bikmukhametov, T., & Jäschke, J. (2020). Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models. *Computers and Chemical Engineering*, *138*, 106834. https://doi.org/10.1016/j.compchemeng.2020.106834

Bollas, G. M., Papadokonstadakis, S., Michalopoulos, J., Arampatzis, G., Lappas, A. A., Vasalos, I. A., & Lygeros, A. (2003). Using hybrid neural networks in scaling up an FCC model from a pilot plant to an industrial unit. *Chemical Engineering and Processing: Process Intensification*, *42*(8–9), 697–713. https://doi.org/10.1016/S0255-2701(02)00206-4

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Camps-Valls, G., Martino, L., Svendsen, D. H., Campos-Taberner, M., Muñoz-Marí, J., Laparra, V., Luengo, D., García-Haro, F. J. (2018). Physics-aware Gaussian processes in remote

sensing. *Applied Soft Computing Journal*, *68*, 69–82. https://doi.org/10.1016/j.asoc.2018.03.021

Chaffart, D., & Ricardez-Sandoval, L. A. (2018). Optimization and control of a thin film growth process: A hybrid first principles/artificial neural network based multiscale modelling approach. *Computers and Chemical Engineering*, *119*, 465–479. https://doi.org/10.1016/j.compchemeng.2018.08.029

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. (2018). Neural ordinary differential equations. In *Advances in Neural Information Processing Systems* (Vol. 2018-Decem, pp. 6571–6583).

Chen, Y., & Ierapetritou, M. (2020). A framework of hybrid model development with identification of plant-model mismatch. *AIChE Journal*, *66*(10), 1–16. https://doi.org/10.1002/aic.16996

Colwell, T., Collet, O., & Downton, J. (2020). Theory-guided data science-based for reservoir characterization. In *1st EAGE Conference on Machine Learning Americas: Applications in the O and G Industry, Challenges and Opportunities* (pp. 1–5). European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.202084013

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1023/A:1022627411411

Cui, Q., & Li, S. (2020). Process monitoring method based on correlation variable classification and vine copula. *Canadian Journal of Chemical Engineering*, *98*(6), 1411–1428. https://doi.org/10.1002/cjce.23702

Dai, X., & Gao, Z. (2013). From model, signal to knowledge: A data-driven perspective of fault detection and diagnosis. *IEEE Transactions on Industrial Informatics*, *9*(4), 2226–2238. https://doi.org/10.1109/TII.2013.2243743

Destro, F., Salmon, A. J., Facco, P., Pantelides, C. C., Bezzo, F., & Barolo, M. (2020). Monitoring a segmented fluid bed dryer by hybrid data-driven/knowledge-driven modeling. In *IFAC-PapersOnLine* (Vol. 53, pp. 11638–11643). https://doi.org/10.1016/j.ifacol.2020.12.646

Downton, J., Collet, O., Colwell, T., 2020. Theory-guided data science-based for Reservoir Characterization, in: First EAGE Conference on Machine Learning in Americas. European Association of Geoscientists & Engineers, pp. 1–5. https://doi.org/10.3997/2214-4609.202084013

Fisher, O. J., Watson, N. J., Escrig, J. E., Witt, R., Porcu, L., Bacon, D., … Gomes, R. L. (2020). Considerations, challenges and opportunities when developing data-driven models for process manufacturing systems. *Computers and Chemical Engineering*, *140*, 106881. https://doi.org/10.1016/j.compchemeng.2020.106881

Gao, Z., Cecati, C., & Ding, S. X. (2015). A survey of fault diagnosis and fault-tolerant techniques-part I: Fault diagnosis with model-based and signal-based approaches. *IEEE Transactions on Industrial Electronics*, *62*(6), 3757–3767. https://doi.org/10.1109/TIE.2015.2417501

Ge, Z. (2017). Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemometrics and Intelligent Laboratory Systems*, *171*, 16–25. https://doi.org/10.1016/j.chemolab.2017.09.021

Georgieva, P., Meireles, M. J., & Feyo de Azevedo, S. (2003). Knowledge-based hybrid modelling of a batch crystallisation when accounting for nucleation, growth and agglomeration phenomena. *Chemical Engineering Science*, *58*(16), 3699–3713. https://doi.org/10.1016/S0009-2509(03)00260-4

Ghosh, D., Moreira, J., & Mhaskar, P. (2021). Model predictive control embedding a parallel hybrid modeling strategy. *Industrial and Engineering Chemistry Research*, *60*(6), 2547–2562. https://doi.org/10.1021/acs.iecr.0c05208

Hendriks, J., Jidling, C., Wills, A., & Schön, T. (2020). Linearly Constrained Neural Networks. arXiv:2002.01600, Retrieved from http://arxiv.org/abs/2002.01600

Hinton, G. (1986). How the backpropagation algorithm works Warm up : a fast matrix-based approach to computing the output. In *Neural Networks and Machine Learning* (pp. 1–25).

Hu, X., Hu, H., Verma, S., & Zhang, Z. L. (2021). Physics-Guided Deep Neural Networks for Power Flow Analysis. *IEEE Transactions on Power Systems*, *36*(3), 2082–2092. https://doi.org/10.1109/TPWRS.2020.3029557

Jack, R. E., Crivelli, C., & Wheatley, T. (2018). Data-Driven Methods to Diversify Knowledge of Human Psychology. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2017.10.002

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168

Kaspar, M. H., & Harmon Ray, W. (1993). Dynamic PLS modelling for process control. *Chemical*

Engineering Science, *48*(20), 3447–3461. https://doi.org/10.1016/0009-2509(93)85001-6

Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, *28*(1), 3–21. https://doi.org/10.1016/0169-7439(95)80036-9

Kreuzinger, T., Bitzer, M., & Marquardt, W. (2008). State estimation of a stratified storage tank. *Control Engineering Practice*, *16*(3), 308–320. https://doi.org/10.1016/j.conengprac.2007.04.013

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote Sensing*, *67*(10), 1137–1142.

Lee, J., & Kang, S. (2007). GA based meta-modeling of BPN architecture for constrained approximate optimization. *International Journal of Solids and Structures*, *44*(18–19), 5980–5993. https://doi.org/10.1016/j.ijsolstr.2007.02.008

Li, G., Qin, S. J., & Yuan, T. (2016). Data-driven root cause diagnosis of faults in process industries. *Chemometrics and Intelligent Laboratory Systems*, *159*, 1–11. https://doi.org/10.1016/j.chemolab.2016.09.006

Li, X., Peng, L., Yao, X., Cui, S., Hu, Y., You, C., & Chi, T. (2017). Long short-term memory neural network for air pollutant concentration predictions: Method development and evaluation. *Environmental Pollution*, *231*, 997–1004. https://doi.org/10.1016/j.envpol.2017.08.114

Lima, P. V., & Saraiva, P. M. (2007). A semi-mechanistic model building framework based on selective and localized model extensions. *Computers and Chemical Engineering*, *31*(4), 361–373. https://doi.org/10.1016/j.compchemeng.2006.07.006

Liu, J., Zheng, W., Zong, Y., Lu, C., & Tang, C. (2020). Cross-corpus speech emotion recognition based on deep domain-adaptive convolutional neural network. *IEICE Transactions on Information and Systems*, *E103D*(2), 459–463. https://doi.org/10.1587/transinf.2019EDL8136

Mavkov, B., Forgione, M., & Piga, D. (2020). Integrated Neural Networks for Nonlinear Continuous-Time System Identification. *IEEE Control Systems Letters*, *4*(4), 851–856. https://doi.org/10.1109/LCSYS.2020.2994806

McBride, K., Sanchez Medina, E. I., & Sundmacher, K. (2020). Hybrid Semi-parametric Modeling

in Separation Processes: A Review. *Chemie-Ingenieur-Technik*. https://doi.org/10.1002/cite.202000025

Muralidhar, N., Islam, M. R., & Marwah, M. (2018). Incorporating Prior Domain Knowledge into Deep Neural Networks. *IEEE Big Data*, 36–45. Retrieved from http://people.cs.vt.edu/naren/papers/PID5657885.pdf

Murphy, C. P., & Kerekes, J. P. (2021). Physics-guided neural network for predicting chemical signatures. *Applied Optics*, *60*(11), 3176. https://doi.org/10.1364/ao.420688

Oliveira, R. (2004). Combining first principles modelling and artificial neural networks: A general framework. In *Computers and Chemical Engineering* (Vol. 28, pp. 755–766). https://doi.org/10.1016/j.compchemeng.2004.02.014

Park, J., Sandberg, I.W., 1991. Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, 3, 246–257. https://doi.org/10.1162/neco.1991.3.2.246

Pawar, S., San, O., Aksoylu, B., Rasheed, A., Kvamsdal, T., 2021. Physics guided machine learning using simplified theories. *Physics of Fluids*, 33, 011701. https://doi.org/10.1063/5.0038929

Psichogios, D. C., & Ungar, L. H. (1992). A hybrid neural network-first principles approach to process modeling. *AIChE Journal, 38(10), 1499–1511*. https://doi.org/10.1002/aic.690381003

Qin, S. J. (1998). Recursive PLS algorithms for adaptive data modeling. *Computers and Chemical Engineering*, *22*(4–5), 503–514. https://doi.org/10.1016/s0098-1354(97)00262-7

Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control,36(2),* 220–234. https://doi.org/10.1016/j.arcontrol.2012.09.004

Rackauckasa, C., Ma, Y., Martensend, J., Warnera, C., Zubove, K., Supekara, R., Skinnera, D., Ramadhana, A., & Edelman, A. (2020) Universal Differential Equations for Scientific Machine Learning. arXiv:2001.04385

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, *348*, 683–693. https://doi.org/10.1016/j.jcp.2017.07.050

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Ren, X., Zhu, K., Cai, T., & Li, S. (2017). Fault Detection and Diagnosis for Nonlinear and Non-Gaussian Processes Based on Copula Subspace Division. *Industrial and Engineering Chemistry Research*, *56*(40), 11545–11564. https://doi.org/10.1021/acs.iecr.7b02419

Safari, S., Shabani, F., & Simon, D. (2014). Multirate multisensor data fusion for linear systems using Kalman filters and a neural network. *Aerospace Science and Technology*, *39*, 465–471. https://doi.org/10.1016/j.ast.2014.06.005

Sansana, J., Joswiak, M. N., Castillo, I., Wang, Z., Rendall, R., Chiang, L. H., & Reis, M. S. (2021). Recent trends on hybrid modeling for Industry 4.0. *Computers and Chemical Engineering*, *151*, 107365. https://doi.org/10.1016/j.compchemeng.2021.107365

Santos, J. E. W., Trierweiler, J. O., & Farenzena, M. (2021). Model Update Based on Transient Measurements for Model Predictive Control and Hybrid Real-Time Optimization. *Industrial and Engineering Chemistry Research*, *60*(7), 3056–3065. https://doi.org/10.1021/acs.iecr.1c00212

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Schuld, M., Sinayskiy, I., & Petruccione, F. (2014). Quantum computing for pattern classification. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8862*, 208–220. https://doi.org/10.1007/978-3-319-13560-1

Simutis, R., & Lübbert, A. (2017). Hybrid approach to state estimation for bioprocess control. *Bioengineering*, *4*(1), 21. https://doi.org/10.3390/bioengineering4010021

Sohlberg, B. (2003). Grey box modelling for model predictive control of a heating process. *Journal of Process Control, 13*(3), 225–238. https://doi.org/10.1016/S0959-1524(02)00030-6

Stefatos, G., & Hamza, A. Ben. (2010). Dynamic independent component analysis approach for fault detection and diagnosis. *Expert Systems with Applications*, *37*(12), 8606–8617. https://doi.org/10.1016/j.eswa.2010.06.101

Stewart, R., & Ermon, S. (2017). Label-free supervision of neural networks with physics and domain knowledge. *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, *1*(1), 2576–2582.

Surtsukov, M.., Neural Ordinary Differential equations, https://msurtsukov.github.io/Neural-ODE/ retreived on Aug 30, 2021.

Svendsen, D. H., Martino, L., Campos-Taberner, M., Garcia-Haro, F. J., & Camps-Valls, G. (2018). Joint Gaussian Processes for Biophysical Parameter Retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, *56*(3), 1718–1727. https://doi.org/10.1109/TGRS.2017.2767205

Svendsen, D. H., Piles, M., Munoz-Mari, J., Luengo, D., Martino, L., & Camps-Valls, G. (2021). Integrating Domain Knowledge in Data-Driven Earth Observation With Process Convolutions. *IEEE Transactions on Geoscience and Remote Sensing*, 1–15. https://doi.org/10.1109/TGRS.2021.3059550

Tan, K. C., & Li, Y. (2002). Grey-box model identification via evolutionary computing. *Control Engineering Practice*, *10*(7), 673–684. https://doi.org/10.1016/S0967-0661(02)00031-X

Thissen, U., Swierenga, H., De Weijer, A. P., Wehrens, R., Melssen, W. J., & Buydens, L. M. C. (2005). Multivariate statistical process control using mixture modelling. *Journal of Chemometrics*, *19*(1), 23–31. https://doi.org/10.1002/cem.903

Thompson, M. L., & Kramer, M. A. (1994). Modeling chemical processes using prior knowledge and neural networks. *AIChE Journal, 40*(8), 1328-1340. https://doi.org/10.1002/aic.690400806

Thornhill, N.F., Patwardhan, S.C., Shah, S.L., 2008. A continuous stirred tank heater simulation model with applications. *Journal of Process Control* 18, 347–360. https://doi.org/10.1016/j.jprocont.2007.07.006

Tidriri, K., Chatti, N., Verron, S., & Tiplica, T. (2016). Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control, 42,* 63–81 . https://doi.org/10.1016/j.arcontrol.2016.09.008

Van Lith, P. F., Betlem, B. H. L., & Roffel, B. (2003). Combining prior knowledge with data driven modeling of a batch distillation column including start-up. *Computers and Chemical Engineering*, *27*(7), 1021–1030. https://doi.org/10.1016/S0098-1354(03)00067-X

Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., Kirsch, B., Walczak, M., Pfrommer, J., Pick, A., Ramamurthy, R., Garcke, J., Bauckhage, C., Schuecker, J. (2021). Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2021.3079836

von Stosch, M., Oliveira, R., Peres, J., & Feyo de Azevedo, S. (2014a). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers and Chemical Engineering*, *60*, 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008

von Stosch, M., Oliveira, R., Peres, J., & Feyo de Azevedo, S. (2014b). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers and Chemical Engineering*, *60*, 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008

Wang, J., Li, Y., Zhao, R., & Gao, R. X. (2020). Physics guided neural network for machining tool wear prediction. *Journal of Manufacturing Systems*, *57*, 298–310. https://doi.org/10.1016/j.jmsy.2020.09.005

Widodo, A., & Yang, B. S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *21*(6), 2560–2574. https://doi.org/10.1016/j.ymssp.2006.12.007

Wu, Z., Rincon, D., & Christofides, P. D. (2020). Process structure-based recurrent neural network modeling for model predictive control of nonlinear processes. *Journal of Process Control*, *89*, 74–84. https://doi.org/10.1016/j.jprocont.2020.03.013

Xiong, Q., & Jutan, A. (2002). Grey-box modelling and control of chemical processes. *Chemical Engineering Science*, *57*(6), 1027–1039. https://doi.org/10.1016/S0009-2509(01)00439-0

Yang, F., Dai, C., Tang, J., Xuan, J., & Cao, J. (2020). A hybrid deep learning and mechanistic kinetics model for the prediction of fluid catalytic cracking performance. *Chemical Engineering Research and Design*, *155*, 202–210. https://doi.org/10.1016/j.cherd.2020.01.013

Yin, S., Ding, S. X., Haghani, A., Hao, H., & Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control*, *22*(9), 1567–1581. https://doi.org/10.1016/j.jprocont.2012.06.009

Yin, S., Ding, S. X., Xie, X., & Luo, H. (2014). A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics, 61(11), 6418–6428*. https://doi.org/10.1109/TIE.2014.2301773

Yu, H., Khan, F., & Garaniya, V. (2015). A probabilistic multivariate method for fault diagnosis of industrial processes. *Chemical Engineering Research and Design*, *104*(3), 306–318. https://doi.org/10.1016/j.cherd.2015.08.026

Yu, J. (2013). A new fault diagnosis method of multimode processes using Bayesian inference based Gaussian mixture contribution decomposition. *Engineering Applications of Artificial Intelligence*, *26*(1), 456–466. https://doi.org/10.1016/j.engappai.2012.09.003

Yu, J., Qin, S.J. (2008). Multimode process monitoring with bayesian inference-based finite Gaussian mixture models. *AIChE Journal*, 54, 1811–1829. https://doi.org/10.1002/aic.11515

Yu, J., & Qin, S. J. (2009). Multiway gaussian mixture model based multiphase batch process monitoring. *Industrial and Engineering Chemistry Research*, *48*(18), 8585–8594. https://doi.org/10.1021/ie900479g

Zhang, S., Bi, K., & Qiu, T. (2020). Bidirectional Recurrent Neural Network-Based Chemical Process Fault Diagnosis. *Industrial and Engineering Chemistry Research*, *59*(2), 824–834. https://doi.org/10.1021/acs.iecr.9b05885

Zhang, Y., & Qin, S. J. (2007). Fault detection of nonlinear processes using multiway kernel independent component analysis. *Industrial and Engineering Chemistry Research*, *46*(23), 7780–7787. https://doi.org/10.1021/ie070381q

Zhang, Y., & Zhang, Y. (2010). Fault detection of non-Gaussian processes based on modified independent component analysis. *Chemical Engineering Science*, *65*(16), 4630–4639. https://doi.org/10.1016/j.ces.2010.05.010

Zhao, H., Sun, S., & Jin, B. (2018). Sequential Fault Diagnosis Based on LSTM Neural Network. *IEEE Access*, *6*, 12929–12939. https://doi.org/10.1109/ACCESS.2018.2794765

Zhou, D., Li, G., & Qin, S. J. (2010). Total projection to latent structures for process monitoring. *AIChE Journal*, *56*(1), 168–178. https://doi.org/10.1002/aic.11977

Chapter 5

A variable mosquito flying optimization based hybrid ANN model for fault detection in process systems

Preface: In this chapter, the performance of a data-driven method has been improved using meta-learning. This is representing the integration of meta-learning of data-driven models. This work can be mapped to the thesis's sub-objective "development of robust data-driven models for the safety of process systems". The content of this chapter has been published as a manuscript in Process Safety Progress, 39 (1), 1-8.

Bibliographic citation and authorship statement: As a first author, I (Md Alauddin) conceptualized the problem, developed codes, prepared the manuscript and responses to the reviewers' comments. The bibliographic citation and contributions of authors and co-authors are presented as follows.

Alauddin, M., Khan, F., Imtiaz, S., & Ahmad, S. (2019). A variable mosquito flying optimization-based hybrid ANN model for the alarm tuning of process fault detection systems. *Process Safety Progress, 39 (1), 1-8.*

Md Alauddin: Formal Analysis, Methodology, Software; Investigation, Validation; Writing - Original Draft; Writing - Review & Editing

Faisal Khan: Conceptualization, Methodology, Writing - Review & Editing; Supervision; Project administration; Funding acquisition

Syed Imtiaz: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition

Salim Ahmed: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition

Abstract: Chemical process systems are becoming extremely complex due to increased automation, heat and mass intensification, and expectation of higher efficiency. Many fault detection and diagnostic methods have been proposed for ensuring safety of processing facilities. However, managing fault detection rate and false alarm rates in the detection and isolation of faults is crucial in complex processing systems. This work presents a new data-driven fault detection model using an artificial neural network and variable mosquito flying optimization technique. The model is based on the optimization of the number of neurons in the hidden layer of the neural network. Subsequently, the model parameters have been tuned using the variable mosquito flying optimization algorithm for maximizing the fault detection rate while minimizing the false alarm rate. The proposed fault detection method has been implemented on the Tennessee Eastman benchmark process. The performance of the proposed model has been evaluated in terms of accuracy, fault detection rate, and false alarm rate against well-known data-driven methods such as principal component analysis, kernel principal component analysis, modified independent component analysis, k nearest neighbors, and support vector machine. The model is observed to be competitive for detecting faults among the test algorithms. This method provides an efficient fault detection tool for complex process systems.

Keywords: *fault detection, process monitoring, neural network, fault detection rate, false alarm rate*

5.1 Introduction:

Alarm management in chemical process industries is becoming increasingly challenging due to the highly correlated multivariate nature of processing systems. A large number of variables and system parameters cause difficulty in the efficient process monitoring of complex industrial systems. Because of the high risk incurred in case of an accident, they are designed for lower missed alarm rates. This results in a myriad of alarms during processing. For instance, an average of 14,250 alarms, with a peak of 26,650 alarms per day were recorded in a European refinery (Ren, Zhu, Cai, & Li, 2017). The exorbitant alarms are usually being ignored by operators. Consequently, alarm flooding is becoming a threat to complex industrial systems. The highly

correlated nature of the multivariate data is another challenge to taking precise mitigative counteractions in real-time. Thus, efficient techniques of process monitoring, fault detection, identification, and diagnosis are needed.

Many fault detection and diagnostic (FDD) methods have been developed over the last four decades. They can broadly be divided into two categories: model-based and data-based. The model-based methods are based on the first principles involving the rigorous development of a process model (Liu, McDermid, & Chen, 2010; Venkatasubramanian, Rengaswamy, Yin, & Kavuri, 2003). These methods have been extensively used in the process systems owing to their robustness and reliability. However, they are not effective in detecting faults of complex processes because of difficulty in capturing the system complexities and nonlinearities in the dynamics. On the other hand, data-driven methods are based on process measurements that do not require a priori quantitative or qualitative knowledge about the process (Isermann, 2005; Staroswiecki, 2000). They can further be divided into qualitative and quantitative classes. Qualitative methods e.g., expert systems (ES) and qualitative trend analysis (QTA) are simple; however, they become complicated for complex process systems. Quantitative methods can be classified into two categories: statistical or non-statistical. Statistical methods such as principal component analysis (PCA) and partial least squares (PLS) use projection of data to a lower-dimensional space for achieving fault detection and diagnosis (Kourti & MacGregor, 1995; Zhou, Li, & Qin, 2010). They can handle large numbers of highly correlated data, measurement errors, and missing data. The dimension reduction feature of these algorithms makes them ideal candidates for combining with other algorithms for handling big data. The major drawback of standard PCA and PLS is that they cannot work with non-Gaussian and nonlinear data (Lawrence, 2005; Liang & Lee, 2013). In addition, Pearson's dependence on the covariance matrix makes PCA sensitive to outliers (Yu,

Khan, & Garaniya, 2016). The kernel principal component analysis (KPCA) transforms nonlinear variables in a higher-dimensional space to render PCA for fault detection (Gharahbagheri, Imtiaz, & Khan, 2017). However, the irreversible nature of kernel function makes them inadequate for fault diagnosis.

Many progressive data-driven methods including artificial neural network (ANN) (Agatonovic-Kustrin & Beresford, 2000), support vector machines (SVM) (Burges, 1998; Cortes & Vapnik, 1995; Widodo & Yang, 2007), classification and regression tree (CART) (Lawrence & Wright, 2001), and Bayes nets (D'Angelo et al., 2014; Verron, Li, & Tiplica, 2010; Zhu, Ge, Song, Zhou, & Chen, 2018) have been studied for detecting abnormalities in process systems.

Artificial neural networks (ANN) have been inspired by biological neural systems. A standard ANN consists of many processors, also called neurons, which generate a sequence of real-valued activations (Schmidhuber, 2015). ANN can assimilate highly complex relationships between several variables by learning dependencies between input and output variables. ANNs are referred to as universal approximators, i.e., they can approximate complex real-valued continuous functions precisely (Park & Sandberg, 1991). Artificial neural networks (ANN) perform well in detecting faults due to its self-learning ability. However, its performance can deteriorate due to intricacy of the error surface. The numerical-based optimization algorithms can lead to suboptimal parameters of the network if it gets trapped at the local optima.

Several hybrid algorithms have been developed to redress this limitation. Nawi, Khan, and Rehman (2013) combined Cuckoo Search (CS) and Levenberg Marquardt algorithm to train a neural network. Genetic algorithm (GA) was also investigated for improving accuracy and speed (Che, Chiang, & Che, 2011; Whitley, Starkweather, & Bogart, 1990). However, most GA-based methods were found to be inadequate for evolving ANN due to two major problems: permutation

and noisy fitness evaluation (Yao & Liu, 1997). Dehuri and Cho (2009) proposed a multi-objective pareto based particle swarm method for simultaneous optimization of architectural complexity and classification accuracy of a neural network. Subudhi and Jena (2011) employed opposition-based differential evolution (ODE) with Levenberg-Marquardt algorithm for training a feed-forward neural network yielding higher accuracy and accelerated convergence rate. Juang (2004) trained a recurrent neural network with a hybrid GA-PSO algorithm using PSO with a master-slave configuration. Subudhi and Jena (2011b) proposed a memetic algorithm-based differential evolution back-propagation (DEBP) for training a multilayer perceptron by exploiting the advantages of both local and global search. Sivagaminathan and Ramakrishnan (2007) exercised a hybrid approach for feature selection using neural networks and ant colony optimization (ACO). They employed ANN for classifying and ACO for evaluating algorithms.

In this work, a population-based metaheuristics, the variable mosquito flying optimization (V-MFO) (Alauddin, 2017) has been used as a hybrid method for improving fault detection performance of the ANN-based model. The model has been developed for maximizing fault detection rate while minimizing the false alarm rate. Section 5.2 provides the methodology of the proposed ANN-VMFO model of detecting faults along with a preliminary introduction to the artificial neural network and variable mosquito flying optimization algorithm. Section 5.3 presents the evaluation of the model for detecting faults of the Tennesee Eastman process followed by a conclusion in Section 5.4.

5.2 The ANN-VMFO model of fault detection

The proposed ANN-VMFO method is a hybrid model based on artificial neural network and a population-based metaheuristics, the variable mosquito flying optimization (V-MFO) algorithm to detect a fault decisively. The model parameters are computed by using multi-objective

optimization based on maximizing fault detection rate while minimizing the false alarm rate. The following sections present an overview of the methodology.

5.2.1: The artificial neural network (ANN)

The artificial neural network detects a fault using its classification and pattern recognition properties. A multilayer neural network comprising an input layer, hidden layer(s) and an output layer can be used for detecting faults of processing systems. The model parameters are iteratively determined using the forward prediction and backward propagation through the network layers. The procedure of fault detection and isolation using the ANN model is based on the selection of an effective optimization algorithm and an optimal number of neurons in the hidden layer of the neural structure. The parameters of the neural network are estimated by iteratively minimizing a multivariate error function.

5.2.2 The variable mosquito flying optimization (V-MFO) algorithm

The V-MFO is a nature-inspired, population-based metaheuristic that mimics the behavior of mosquitoes to find holes or irregularities in a mosquito net (Alauddin, 2017). Mosquitoes explore mosquito nets for finding holes or irregularities through which they get to their prey (Fig. 5.1). The search comprises to and fro flying and sliding motion, i.e., they fly to a point on the net and walk in any direction for a few steps as depicted in Fig 5.1a. The movements of the mosquitoes over a net are described in Fig. 5.1b-d. Fig. 5.1b depicts that a mosquito flying from point 'A', reaches point 'B' by a flying motion (solid line) followed by a sliding motion (dotted line) to reach point 'C' or 'C''. The region explored by a flight or slide is governed by numerous factors such as the type and size of mosquitoes, attributes of the net, nature of prey, time of the attack, presence of external force, and surrounding characteristics. Moreover, the distance covered in each flight and slide may or may not be equal (Fig. 5.1c-d). Fig. 5.1d shows that the distance covered in the

143

sliding cycle which starts at point B may be ended by point C, C', or C''. After getting access to their prey they explore a suitable suction point on the body of the victim.
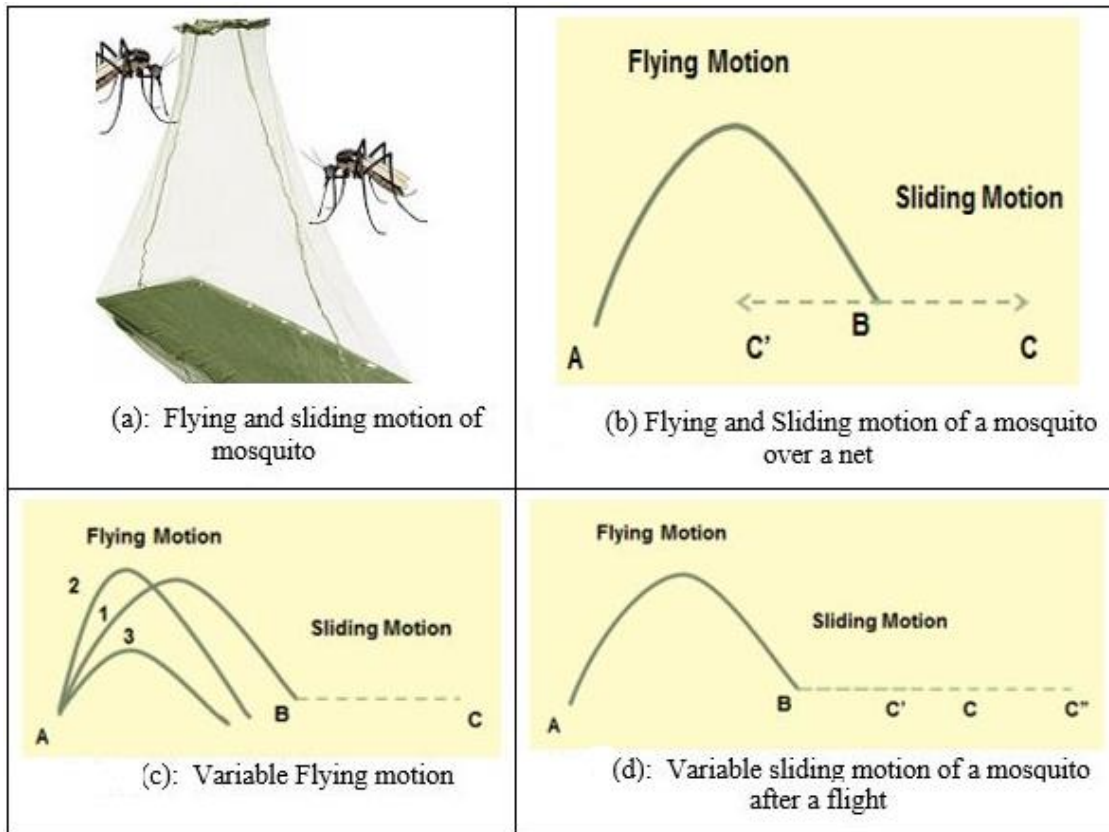
Fig. 5.1: Movements of mosquitoes over a net for searching of prey (Alauddin, 2017)

The procedure for the variable mosquito flying optimization (V-MFO) algorithm is as follows-

STEP 1   Initiate the population of particles for each variable.

STEP 2   Evaluate fitness of the objective function.

STEP 3   Apply flying motion to the particles: Par= Phi1/Phi1*par

      (It can apply in both directions: for forward flight multiply by phi1 (golden ratio, 1.618)

                    for the backward flight divide by phi1 (the golden ratio, 1.618)

STEP 4   Evaluate new particles

STEP 5   Modify fitness function if new fitness is better

STEP 6   Apply sliding motion: Par= par*(1 $\pm c_1$)

$c_1 = 0.1$ is a sliding constant

Par= par*(1 +c_1): for the forward sliding

Par= par*(1 -c_1): for the backward sliding

STEP 7   Evaluate the new fitness

STEP 8   Modify the fitness function if new fitness is better

STEP 9   Apply precision movement   Par= par*(1 ± c_2)

$c_2 = 0.01$ is a precision constant

Par= par*(1 +c_2): for the forward movement

Par= par*(1 -c_2): for the backward movement

STEP 10   Evaluate the new fitness

STEP 11   Modify the fitness function if new fitness is better

STEP 12   Decrement iteration counter and evaluate the deviation. If the iteration counter is not

zero OR the deviation is less than DEVMIN, then go to STEP 3

STEP 13 STOP

5.2.3 Fault detection using the ANN-VMFO hybrid model

The flowchart of the ANN-VMFO hybrid network for fault detection is shown in Fig. 5.2. The

offline refers to the data used for the model development while the online stands for the real-time

process monitoring data. The offline data is normalized and divided into three sets: training

validation and testing. As the number of iterations increases, the training performance improves

because it tries to fit each data accurately. This can result in overfitting leading to increased

complexity and compromised generalization, i.e., the model will not perform well with new or

unseen data. This can be handled using validation that penalizes some of the weights leading to a

more generalized structure. The optimum number of hidden nodes and layers depends on

input/output sizes, training and test data sizes, and the characteristics of the problem (Kiranyaz,

Ince, Yildirim, & Gabbouj, 2009). The higher the number of neurons in the neural structure more complex will be the error function. The performance of the optimization algorithms is subjected to the type and complexity of the problem. Having been iterative, the algorithms tend to get trapped at the local optima yielding suboptimal results. The present work addresses this by using a hybrid method of deterministic and population-based metaheuristics to find global optima. This is achieved by tuning the parameters using the variable mosquito flying optimization algorithm for maximizing fault detection rate while minimizing the false alarm rate.
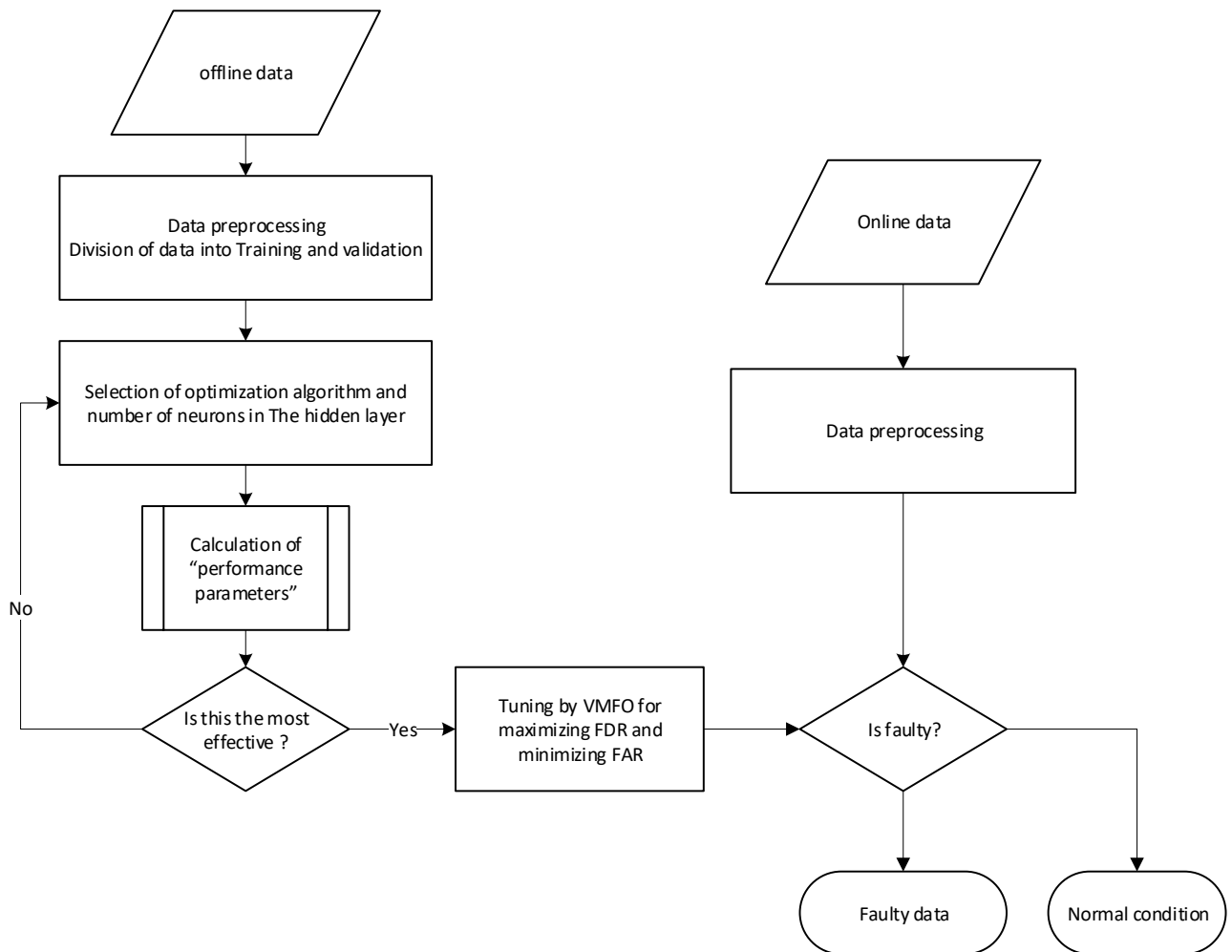
Fig. 5.2: The flowchart of the ANN-VMFO model for process fault detection

5.3 Results and Discussion

The proposed fault detection model has been evaluated on the Tennessee Eastman chemical process. These evaluations were based on the classifier properties such as accuracy, fault detection rate (FDR), and false alarm rate (FAR). Accuracy is the fraction of correct prediction by the model. It represents the overall effectiveness of a classifier. The FDR announces the fraction of the faults rightly detected by a classifier whereas FAR indicates the fraction of the alarms turned out false. The accuracy, fault detection rate, false alarm rate, and missed detection rate can be calculated using Eq. 5.1, Eq. 5.2, Eq. 5.3, and Eq. 5.4, respectively (Chen, Tiňo, & Yao, 2014; Jaffel, Taouali, Elaissi, & Messaoud, 2014; Zhao et al., 2017).

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{5.1}$$

$$FDR = \frac{Fault\ detected}{Total\ number\ of\ faults} = \frac{TP}{TP+FP} \tag{5.2}$$

$$FAR = \frac{False\ alarm}{Total\ number\ of\ alarms} = \frac{FN}{TP+FN} \tag{5.3}$$

$$MDR = \frac{Missed\ alarm}{Total\ number\ of\ expected\ alarms}$$

$$= \frac{Fault\ signal\ classified\ as\ normal}{Total\ number\ of\ faults} \tag{5.4}$$

Where,   TP= true positive (fault signal correctly diagnosed as fault),

TN= true negative (normal signal correctly diagnosed as normal),

FP= false positive (fault signal incorrectly identified as normal),

and FN= false negative (normal signal incorrectly identified as fault).

Accuracy and fault detection rate should be higher whereas missed detection rate and the false alarm rate should be lower for an effective detection system.

5.3.1 *Fault detection of the Tennessee Eastman chemical process*

The Tennessee Eastman chemical process comprises five major operating units: an exothermic two-phase reactor, a product condenser, a vapor-liquid flash separator, a recycle compressor, and a reboiled product stripper. The reactor is charged with three gaseous reactants which form a liquid product as a result of a catalyzed chemical reaction. The product stream is fed to a condenser followed by a vapor-liquid separator. The non-condensed product is recycled back to the reactor a centrifugal compressor whereas the condensed output is fed to a stripper. The final product stream from the stripper's bottom is pumped downstream for further processing (Downs & Vogel, 1993). The Tennessee Eastman process comprises 41 measured and 12 manipulated variables. Among the measured variables, 22 variables are continuous process variables and 19 variables are related to composition measurements. The simulation was run on the data of the Harvard database (https://doi.org/10.7910/DVN/6C3JR1). The subsequent sections describe the performance of the proposed algorithm for fault detection and fault isolation on the TE process.

The model was trained and tested on 20 different types of faults of the TE process (Table 5.1). Data were divided in two classes; normal and faulty. The optimal neural structure was determined by minimizing the cross-entropy function by varying the number of neurons for the distinct algorithms. The cross-entropy for each pair of output/target elements is calculated as follows; CE = -t .* log(y), where, t and y are the targets and the output respectively. The aggregate cross-entropy performance is the mean of the individual values. It was observed that the Bayesian regularization algorithm performing better among the selected algorithms. Thus, an optimal neural network structure was specified with the Bayesian regularization algorithms for the backpropagation and the number of neurons n=20 in the hidden layer.

Table 5.1: The distinct faults of the TE Processes

Faults	Description	Type of Faults
IDV-1	A/C feed ratio, B composition constant (Stream 4)	Step
IDV-2	B composition, A/C ratio constant (Stream 4)	Step
IDV-3	D feed temperature (Stream 2)	Step
IDV-4	Reactor cooling water inlet temperature	Step
IDV-5	Condenser cooling water inlet temperature	Step
IDV-6	A feed loss (Stream 1)	Step
IDV-7	C header pressure loss-reduced availablity (Stream 4)	Step
IDV-8	A, B, C feed composition (Stream 4)	Random Variation
IDV-9	D feed temperature (Stream 2)	Random Variation
IDV-10	C feed temperature (Stream 4)	Random Variation
IDV-11	Reactor cooling water inlet temperature	Random Variation
IDV-12	Condenser cooling water inlet temperature	Random Variation
IDV(13)	Reaction kinetics	Slow Drift
IDV(14)	Reactor cooling water valve	Sticking
IDV(15)	Condenser cooling water valve	Sticking
IDV(16)	Unknown	
IDV(17)	Unknown	
IDV(18)	Unknown	
IDV(19)	Unknown	
IDV(20)	Unknown	

The proposed technique has been compared with well-known data-based methods such as nearest neighbors (KNN), discriminant analysis (DA), support vector machines (SVM), and ensemble learning (bagged tree and boosted tree). The hybrid ANN-VMFO method registered the highest accuracy (97.89%) followed by the standard ANN (97.85%), and the bagged tree (94.16%). The fine KNN (89.37%), fine tree (86.71%) and SVM (84.43%) and boosted tree (81.74%) are other effective algorithms for detecting faults of the Tennessee Eastman chemical process (Fig. 5.3A).
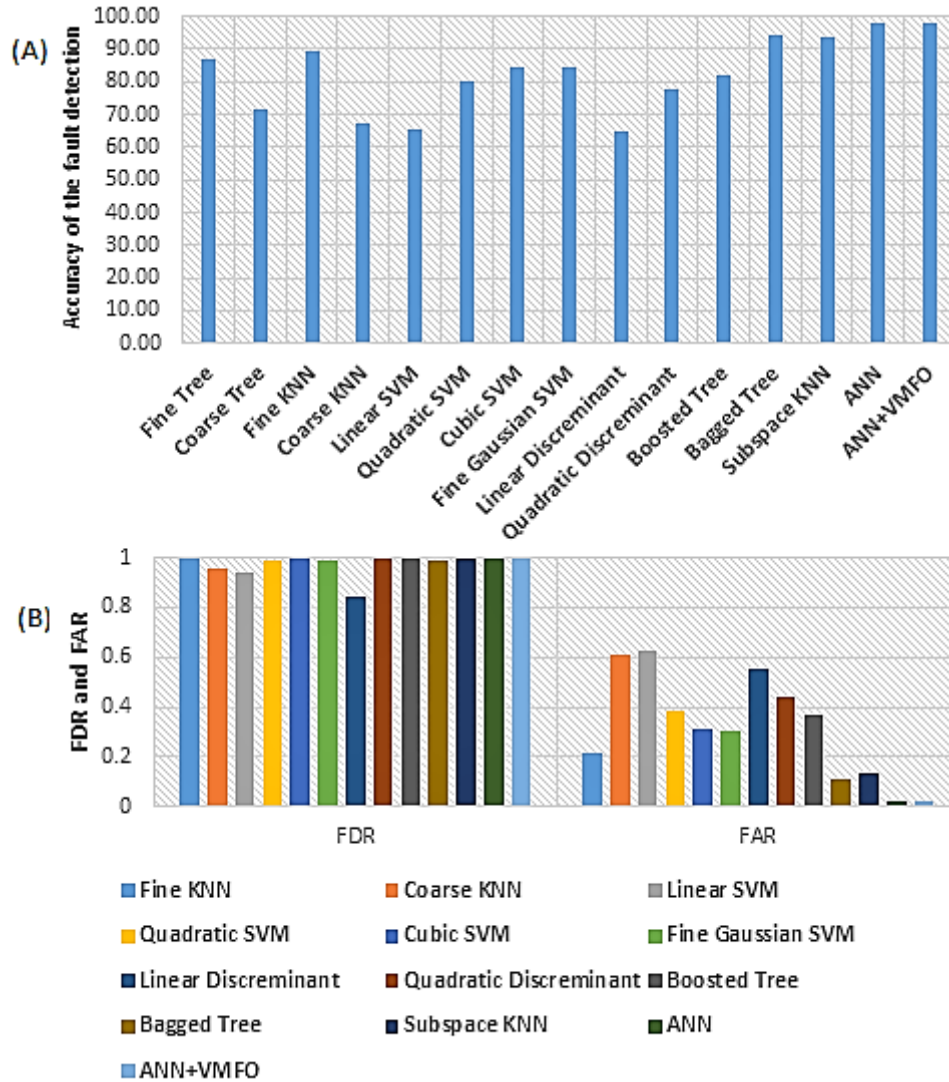
Fig. 5.3: Comparative performance of detection system in terms of A: accuracy, B: Fault detection rate (FDR), and false alarm rate (FAR)

Fig. 5.3B demonstrates fault detection rate (FDR) and the false alarm rate (FAR) of selected methods on the case study. Most of these algorithms perform well with more than 95% fault detection rate. The fine KNN subspace KNN, quadratic discriminant and boosted tree records 100% FDR. Nonetheless, they result in a higher false alarm rate (21.26%, 44.39%, and 13.1% respectively) which is undesirable. However, the proposed ANN-VMFO model yield higher FDR (0.9967) and a lower FAR (0.0184).

The performance of the fault detection system was improved by employing the variable mosquito flying optimization (VMFO) algorithm for tuning the neural network parameters. The detailed procedure of the parameter tuning using the ANN-VMFO algorithm has been presented in Fig 5.4. The parameters are determined by minimizing the error subjected to maximizing fault detection rate (FDR) while minimizing the false alarm rate (FAR).

The confusion matrix for the fault detection for the standard ANN and the proposed ANN-VMFO hybrid algorithm have been presented in Fig. 5.5A and Fig. 5.5B respectively. It can be observed that the standard ANN accurately detects 9963 faults out of 10000 faults whereas the ANN-VMFO model can detect 9966 faults at a similar condition. The comparative analysis of the standard ANN and the ANN-VMFO model has been presented in Table 5.2. The proposed ANN-VMFO model does not yield a significant improvement in terms of accuracy and the fault detection rate. Nonetheless, it resulted in a 0.6 % improvement in the false alarm rate (FAR) and an 8% improvement in missed detection rate (MDR) compared to the simple ANN.

Table 5.2: The improved network performance by using hybrid ANN-VMFO model

Performance Factors	ANN	ANN-VMFO	% Improvement
Accuracy	0.9785	0.9789	0.04
Fault detection rate	0.9963	0.9966	0.03
False alarm rate	0.0185	0.0184	0.55
Missed detection rate	0.0037	0.0034	8.82

```
Results from ANN
        ANN Parameters
                Weights:
                        Hidden Layer weights-W11(SXR)
                        Output layer weights- w21 (YXS)
                Biases:
                        Hidden layer biases: b11 (Sx1)
                        output layer biases: b21 (Yx1)


                Fault detection rate:  F1
                False Alarm Rate:      F2


Tuning of parameters of ANN using VMFO for better results
        Weights:
                        Hidden Layer weights-W11'(SXR)
                        Output layer weights- w21' (SXQ)
                Biases:
                        Hidden layer biases: b11' (Sx1)
                        output layer biases: b21' (Yx1)


                Fault detection rate:   F1'
                False Alarm Rate:       F2'


The objective is to reduce the misclassification using multi-objective optimization of maximization of
fault detection rate and minimization of false alarm rate


Such that:       F1'> F1 AND F2'< F2
                 OR  F1'>F1 AND F2'< =F2
                 OR F1'>= F1 AND F2'< F2
                 OR F1'>= F1 AND F2'<= F2
```

Fig. 5.4: Tuning of model parameters using the hybrid ANN-VMFO model

(A)				(B)		
Normal	312	188		Normal	313	187
Fault	37	9963		Fault	34	9966
	Normal	Fault			Normal	Fault

Observed (A) / *Predicted* ; *Observed* (B) / *Predicted*

Fig. 5.5: Confusion matrix for the fault detection system of TE process: (A) using simple ANN (B) using ANN-VMFO hybrid algorithms

152

5.4 Conclusions

This paper presented a hybrid model for detecting faults of process systems based on artificial neural network (ANN) and variable mosquito flying optimization (V-MFO) technique. The model was evaluated on the benchmark Tennessee Eastman chemical process. The model was trained and tested with 10500 data with 500 sample data for each normal and 20 different types of faults. The proposed technique was compared with prominent data-based methods. The proposed ANN-VMFO hybrid method registered an optimum accuracy (97.89%) followed by the simple ANN (97.85%) and bagged tree (94.16%). The fine KNN (89.37%), fine tree (86.71%), SVM (84.43%) and boosted tree (81.74%) were other effective algorithms for detecting faults of the Tennessee Eastman chemical process. The parameter tuning using the variable mosquito flying optimization (V-MFO) resulted in slightly improved accuracy, fault detection rate, and false alarm rate whereas, it demonstrated significant improvement in missed detection rate compared to the simple ANN. The proposed ANN-VMFO model could be an efficient fault detection tool for complex process systems.

List of symbols and abbreviations

Symbols	Meanings
A, B, C, C', C"	position in space
ACO	ant colony optimization
ANN	artificial neural network
b11	bias of the hidden layer
b12	bias of the output layer
c_1	sliding constant
c_2	precision constant
CART	classification and regression tree
CBR	case-based reasoning
CE	cross-entropy
CS	cuckoo search

DA	discriminant analysis
DEBP	differential evolution back-propagation
ES	expert systems
F1, F1'	fault detection rate
F2, F2'	false alarm rate
FAR	false alarm rate
FDD	fault detection and diagnosis
FDR	fault detection rate
FN	true negative
FP	true positive
GA	genetic algorithm
GMM	Gaussian mixture model
ICA	independent component analysis
IDV-1 to 20	faults of the simulated TE Process
KNN	k nearest neighbors
MDR	missed detection rate
PCA	principle component analysis
Phi1	golden ratio
PLS	partial least square
QTA	qualitative trend analysis
SVM	Support vector machine
TN	true negative
TP	true positive
V-MFO	variable mosquito flying optimization
W11	weight of the hidden layer
W12	weight of the output layer
XMEAS	input variable in the TE Process

References

Agatonovic-Kustrin, S., & Beresford, R. (2000). Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *Journal of Pharmaceutical and Biomedical Analysis*, *22*(5), 717–727. https://doi.org/10.1016/S0731-7085(99)00272-1

Alauddin, M. (2017). V-MFO: Variable Flight Mosquito Flying Optimization. In *Applications of Soft Computing for the Web* (pp. 271–283). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-10-7098-3_15

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167. https://doi.org/10.1023/A:1009715923555

Che, Z. G., Chiang, T. A., & Che, Z. H. (2011). Feed-forward neural networks training: A comparison between genetic algorithm and back-propagation learning algorithm. *International Journal of Innovative Computing, Information and Control*, *7*(10), 5839–5850.

Chen, H., Tiňo, P., & Yao, X. (2014). Cognitive fault diagnosis in Tennessee Eastman Process using learning in the model space. *Computers and Chemical Engineering*, *67*, 33–42. https://doi.org/10.1016/j.compchemeng.2014.03.015

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning, 20, 273-297*. https://doi.org/10.1023/A:1022627411411

D'Angelo, M. F. S. V., Palhares, R. M., Cosme, L. B., Aguiar, L. A., Fonseca, F. S., & Caminhas, W. M. (2014). Fault detection in dynamic systems by a Fuzzy/Bayesian network formulation. *Applied Soft Computing Journal*, *21*, 647–653. https://doi.org/10.1016/j.asoc.2014.04.007

Dehuri, S., & Cho, S. B. (2009). Multi-criterion Pareto based particle swarm optimized polynomial neural network for classification: A review and state-of-the-art. *Computer Science Review*, *3*(1), 19–40. https://doi.org/10.1016/j.cosrev.2008.11.002

Downs, J. J., & Vogel, E. F. (1993). A plant-wide industrial process control problem. *Computers and Chemical Engineering*, *17*(3), 245–255. https://doi.org/10.1016/0098-1354(93)80018-I

Gharahbagheri, H., Imtiaz, S. A., & Khan, F. (2017). Root Cause Diagnosis of Process Fault Using KPCA and Bayesian Network. *Industrial and Engineering Chemistry Research*, *56*(8), 2054–2070. https://doi.org/10.1021/acs.iecr.6b01916

Isermann, R. (2005). Model-based fault-detection and diagnosis - Status and applications. *Annual Reviews in Control*, *29*(1), 71–85. https://doi.org/10.1016/j.arcontrol.2004.12.002

Jaffel, I., Taouali, O., Elaissi, E., & Messaoud, H. (2014). A new online fault detection method based on PCA technique. *IMA Journal of Mathematical Control and Information*, *31*(4), 487–499. https://doi.org/10.1093/imamci/dnt025

Juang, C. F. (2004a). A Hybrid of Genetic Algorithm and Particle Swarm Optimization for Recurrent Network Design. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(2), 997–1006. https://doi.org/10.1109/TSMCB.2003.818557

Juang, C. F. (2004b). A Hybrid of Genetic Algorithm and Particle Swarm Optimization for Recurrent Network Design. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *34*(2), 997–1006. https://doi.org/10.1109/TSMCB.2003.818557

Kiranyaz, S., Ince, T., Yildirim, A., & Gabbouj, M. (2009). Evolutionary artificial neural networks

by multi-dimensional particle swarm optimization. *Neural Networks*, *22*(10), 1448–1462. https://doi.org/10.1016/j.neunet.2009.05.013

Kourti, T., & MacGregor, J. F. (1995). Process analysis, monitoring and diagnosis, using multivariate projection methods. *Chemometrics and Intelligent Laboratory Systems*, *28*(1), 3–21. https://doi.org/10.1016/0169-7439(95)80036-9

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research*, *6*, 1783–1816.

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote Sensing*, *67*(10), 1137–1142.

Liang, Z., & Lee, Y. (2013). Eigen-analysis of nonlinear PCA with polynomial kernels. *Statistical Analysis and Data Mining*, *6*(6), 529–544. https://doi.org/10.1002/sam.11211

Liu, S., McDermid, J. A., & Chen, Y. (2010). A rigorous method for inspection of model-Based formal specifications. *IEEE Transactions on Reliability*, *59*(4), 667–684. https://doi.org/10.1109/TR.2010.2085571

Nawi, N. M., Khan, A., & Rehman, M. Z. (2013). A New Levenberg Marquardt based Back Propagation Algorithm Trained with Cuckoo Search. *Procedia Technology*, *11*(Iceei), 18–23. https://doi.org/10.1016/j.protcy.2013.12.157

Park, J., & Sandberg, I. W. (1991). Universal Approximation Using Radial-Basis-Function Networks. *Neural Computation*, *3*(2), 246–257. https://doi.org/10.1162/neco.1991.3.2.246

Ren, X., Zhu, K., Cai, T., & Li, S. (2017). Fault Detection and Diagnosis for Nonlinear and Non-Gaussian Processes Based on Copula Subspace Division. *Industrial and Engineering Chemistry Research*, *56*(40), 11545–11564. https://doi.org/10.1021/acs.iecr.7b02419

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Sivagaminathan, R. K., & Ramakrishnan, S. (2007). A hybrid approach for feature subset selection using neural networks and ant colony optimization. *Expert Systems with Applications*, *33*(1), 49–60. https://doi.org/10.1016/j.eswa.2006.04.010

Staroswiecki, M. (2000). Quantitative and qualitative models for fault detection and isolation. *Mechanical Systems and Signal Processing*, *14*(3), 301–325. https://doi.org/10.1006/mssp.2000.1293

Subudhi, B., & Jena, D. (2011). Nonlinear system identification using memetic differential evolution trained neural networks. *Neurocomputing*, *74*(10), 1696–1709. https://doi.org/10.1016/j.neucom.2011.02.006

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis. *Computers & Chemical Engineering*, *27*(3), 293–311. https://doi.org/10.1016/S0098-1354(02)00160-6

Verron, S., Li, J., & Tiplica, T. (2010). Fault detection and isolation of faults in a multivariate process with Bayesian network. *Journal of Process Control*, *20*(8), 902–911. https://doi.org/10.1016/j.jprocont.2010.06.001

Whitley, D., Starkweather, T., & Bogart, C. (1990). Genetic algorithms and neural networks: optimizing connections and connectivity. *Parallel Computing*, *14*(3), 347–361. https://doi.org/10.1016/0167-8191(90)90086-O

Widodo, A., & Yang, B. S. (2007). Support vector machine in machine condition monitoring and fault diagnosis. *Mechanical Systems and Signal Processing*, *21*(6), 2560–2574. https://doi.org/10.1016/j.ymssp.2006.12.007

Yao, X., & Liu, Y. (1997). A new evolutionary system for evolving artificial neural networks. *IEEE Transactions on Neural Networks*, *8*(3), 694–713. https://doi.org/10.1109/72.572107

Yu, H., Khan, F., & Garaniya, V. (2016). An Alternative Formulation of PCA for Process Monitoring Using Distance Correlation. *Industrial and Engineering Chemistry Research*, *55*(3), 656–669. https://doi.org/10.1021/acs.iecr.5b03397

Zhao, H., Liu, J., Dong, W., Sun, X., & Ji, Y. (2017). An improved case-based reasoning method and its application on fault diagnosis of Tennessee Eastman process. *Neurocomputing*, *249*, 266–276. https://doi.org/10.1016/j.neucom.2017.04.022

Zhou, D., Li, G., & Qin, S. J. (2010). Total projection to latent structures for process monitoring. *AIChE Journal*, *56*(1), 168–178. https://doi.org/10.1002/aic.11977

Zhu, J., Ge, Z., Song, Z., Zhou, L., & Chen, G. (2018). Large-scale plant-wide process modeling and hierarchical monitoring: A distributed Bayesian network approach. *Journal of Process Control*, *65*, 91–106. https://doi.org/10.1016/j.jprocont.2017.08.011

Chapter 6

How Can Process Safety and a Risk Management Approach Guide Pandemic Risk Management?

Preface: *Although the Coronavirus disease of 2019 did not result from process operations, energy industries were one the hardest hit sectors due to the present pandemic. The COVID-19 caused severe disruption of processing operations leading to a historic collapse in demand and price. It also affected employment safety and operability of process industries and allied sectors. This chapter presents an advanced semi-mechanistic model that can capture temporal variability in assessing pandemic risk. It also accounts for uncertainty in the parameters in various controlling and mitigative actions. The last section discusses how process safety practices are applicable to pandemic risk management. This work can be mapped to two sub-objectives of the thesis, devising semi-mechanistic models for assessing pandemic risk and to establish synergy between process safety and pandemic risk management. This chapter has been published in Journal of Loss Prevention in Process Industries, 68, 104310.*

Abstract: The coronavirus disease of 2019 brought the world to a halt in March 2020. Various risk assessment and management approaches are being explored worldwide for more informed decision-making. This work adopts a semi-mechanistic model and a process safety framework for managing the risk of the current pandemic. A parameter tweaking and an artificial neural network-based parameter learning of the susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) models have been devised to forecast the dynamic risk of infectious diseases. The randomness of the model parameters has been captured using Monte Carlo simulation. The proposed models have been studied on assessing the infection risk of the COVID-19 at four locations: Italy, Germany, Ontario, and British Columbia. The learning-based approach resulted in better outcomes among the models tested in the present study. A layer of protection analysis for distinct measures of pandemic risk management has been proposed as well. We also quantified risk under enforcement and release of distinct risk-reducing measures. The risk profiles suggest that a stage-wise releasing scenario is the optimal approach to getting back to normal. The case study provides valuable insights to practitioners in the health sector and the process industries to implement advanced strategies for risk assessment and management. This work explores a synergy between process safety and epidemiology to better understand, analyze, and manage the risk using advanced mathematical models, management tools, and, more importantly, the lessons learned from crises.

Keywords: Risk, process monitoring, neural network, pandemic, non-pharmaceutical interventions; layers of protection analysis.

6.1. Introduction
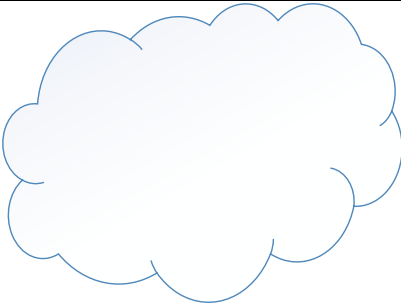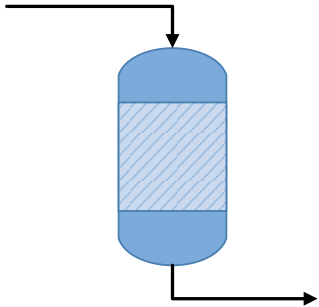
The coronavirus disease (COVID-19) has been declared a global pandemic by the World Health Organization (WHO). The high infection rate of the coronavirus and the unavailability of a vaccine have led to an unprecedented scenario. Countless numbers of people were deprived of proper medical care due to the saturation of health care facilities at many places. More than 28 million infected cases and over nine hundred thousand mortalities due to the outbreak have been reported to date (Worldometer, Sep 09, 2020).
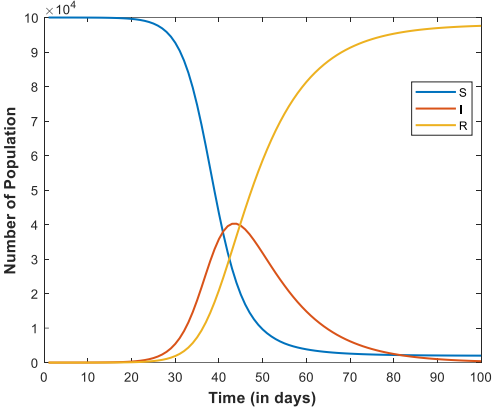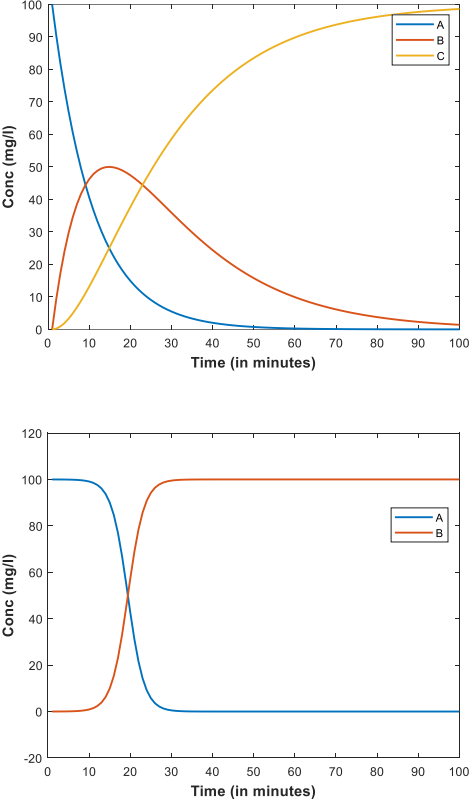
In epidemiological studies, mechanistic models have been widely used for pandemic risk management. Kermack and McKendrick (1927) developed the SIR (susceptible, infected, recovered) model which was subsequently revised by (Anderson & May, 1979; Hethcote, 1976; Hiorns & MacDonald, 1982). The SIR model assumes that the infected hosts instantaneously become infectious after being exposed to an infected carrier. However, infectious diseases usually have latency i.e. incubation period of the virus before an infected agent truly becomes infectious. This was addressed by the SEIR (susceptible, exposed, infected, recovered) model by including an exposed (E) compartment to the SIR model. Many models with varying population sizes and distinct vaccination strategies towards susceptible individuals have been developed to date (Busenberg & Driessche, 1990; Li, Graef, Wang, & Karsai, 1999; Martcheva & Castillo-Chavez, 2003; Sun & Hsieh, 2010). The basic SIR and SEIR models do not reflect hospitalization and quarantine effects that effectively lower the cumulative infection cases. The impacts of quarantine, isolation, and other nonpharmaceutical interventions (NPIs) to restrain infectious diseases have been presented by many studies including (Hethcote, Zhien & Shengbing, 2002; Hollingsworth, 2009; Lipsitch et al., 2003; Safi & Gumel, 2010). The basic and the modified SEIR models have been used to investigate the spread of several diseases such as SARS (2003), Influenza (2009), Ebola (2014), MERS (2015), and Zika (2016) (Zhang et al., 2017).

Despite different disciplines, process safety and epidemiology share many similarities in modeling the dynamic behavior of underlying processes. Compartmental models have been used in modeling pandemic diseases and numerous chemical processing systems; e.g., continuous stirred tank reactor, fluidized bed reactor, and bubble column for studying kinetics, velocity distribution, energy dissipation rate, crystal size distribution, and turbulence (Alvarado, Vedantam, Goethals, & Nopens, 2012; Bermingham, Kramer, & Rosmalen, 1998; Cui, Van Der Lans, Noorman, &

Luyben, 1996; Iliuta, Larachi, Anfray, Dromard, & Schweich, 2007; Vrábel, Van Der Lans, Cui, & Luyben, 1999; Zhao, Buffo, Alopaeus, Han, & Louhi-Kultanen, 2017). There is a strong resemblance between the disease dynamics of epidemiological model(s) and the kinetic model(s) of chemical reactors, especially the continuous stirred tank reactor (CSTR). Table 6.1 presents similarities between the SIR epidemiological model and the reaction kinetics model of a CSTR. Social structural complexity, distinct individual practices, and economic disparity, however, complicate epidemic modeling. Randomness in the parameters (e.g., incubation, infection, and recovery periods) also makes epidemic modeling difficult compared to the reactor models. Demographics and chronic health conditions significantly affect susceptibility in epidemic modeling.

Table 6.1: Similarities between epidemiological and reactor kinetics model.

	SIR epidemiological model	Reactor kinetic model
System	 Geographical Location	 Chemical Reactor
Propagation/ Reaction model	$S \rightarrow I \rightarrow R$	Series reaction : $A \overset{k_1}{\rightarrow} B \overset{k_2}{\rightarrow} C$ Auto-catalytic reaction: $A + B \overset{k}{\rightarrow} 2\,B$
Model equations	$\dfrac{dS}{dt} = -\dfrac{aS(t)I(t)}{N}$ $\dfrac{dI}{dt} = \dfrac{aS(t)I(t)}{N} - bI(t)$ $\dfrac{dR}{dt} = bI\,(t)$	Series Reaction $\dfrac{dC_A}{dt} = -k_1 C_A(t)$ $\dfrac{dC_B}{dt} = k_1 C_A(t) - k_2 C_B(t)$ $\dfrac{dC_C}{dt} = k_2 C_B(t)$ Autocatalytic Reaction (in its simplest form) $\dfrac{dC_B}{dt} = kC_A(t)C_B(t)$

Response Curves		
Principles of conservation	Follows conservation principles; i.e., within a geographical area, the total number of people in all of the compartments remains constant. $$N = S + I + R$$	Follows the conservation of mass principles; i.e., the total mass of all of the species remains constant. $$M = M_A + M_B + M_C$$
Rate of spread/ transformations	The contagion rate is the determining factor for the states of the epidemic. It depends on the basic reproduction numbers of the epidemic, government regulations (e.g., limiting gathering size, closure of nonessential business and schools, emergency lockdown), and personal hygiene measures such as wearing masks in public places, frequent washing of hands, and social distancing.	The reaction rate governs transformations in chemical reactions. Rates of chemical transformation are often affected by the rates of other processes such as heat or mass transfer, the presence of a catalyst, and species concentration, dispersion, segregation, and mixing.

Different non-pharmaceutical interventions (NPIs) are advocated by healthcare authorities to control the spread of a pandemic. Social distancing, frequent hand washing, wearing a mask, and practicing good hygiene are effective NPIs for slowing down the spread (Ferguson et al., 2020; Davies et al., 2020).

Similar to epidemiology, risk minimization is a central aspect of process safety systems (Khan & Abbasi, 1998). The methodology to prevent, control, and mitigate infection is analogous to the hazard control of process safety frameworks of industrial systems. Over the years, different qualitative and quantitative safety management systems have been devised for reducing risk in process industries. The layer of protection analysis (LOPA) and inherently safer design (ISD) are two of the most promising risk assessment and management tools (Khan et al., 2015). LOPA with inherent safety considerations provides better insights for decision-making. Some earlier approaches to implementing LOPA for improving safety management are discussed in (Dowell, 1998; Srinivasan & Natarajan, 2012). Gowland (2006) explained the principle of LOPA and its applicability for accidental risk assessment. An application of LOPA to estimate the risk due to reactive chemicals is presented in (Wei, Rogers, & Mannan, 2008) and an improved version of LOPA called ExSys-LOPA is proposed in (Markowski & Mannan, 2010). At the core of the LOPA framework, preventive measures are used to avoid a probable abnormal event. In the next layer, the control system is in place to counteract and slow down the escalation of the abnormalities. Following that is the emergency safety layer to restrain and nullify the impact of hazard when the control system fails. A comprehensive review of the existing literature on process safety and risk assessment is reported in (Khan, Rathnayaka, & Ahmed, 2015).

The motivating factors for this work are to devise semi-mechanistic models for assessing pandemic risk and to establish synergy between process safety and pandemic risk management. We have proposed a layer of protection analysis (LOPA) for epidemic management. The performance of the safety management framework has been quantified in terms of the reduction of pandemic risk when safety layers are in effect.

We have proposed a parameter tweaking and an artificial neural network-based parameter learning method for computing adaptive parameters of the susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) model. The proposed semi-mechanistic SEIQRD model has been employed to predict the probability of fatalities and severity of consequences, thus enabling risk evaluation. The model has also been used to assess risk in enforcing and releasing distinct intervention strategies. On the other hand, LOPA and inherent safety methods were used for risk management. By integrating the two approaches, a framework was developed for risk assessment and management of a pandemic, including mitigating, suppressing, and releasing factors.

The paper is organized as follows: Section 2 describes the mathematical models used to assess risk. Section 3 discusses the outcome of the risk calculations from the proposed models for different geographical locations. Section 3 also discusses the impact of the LOPA mitigation factors on risk. Finally, risk profiles for the releasing scenarios are presented. Section 4 gives the conclusion and future research directions.

6.2 The Mathematical Model

This section describes the mathematical models to predict the spread of epidemic diseases. It begins by introducing the mechanistic model of the epidemic spread followed by the approaches to solving the model using parameter tweaking and the ANN-based parameter learning model.

6.2.1 The SEIQRD epidemic model

Many studies reveal that people may be more contagious around the time of symptom onset than the diseased one (World Health Organization, 2020). Numerous reports highlighted that the pre-symptomatic (infectious before the symptom onset) and asymptomatic (does not develop symptom) are the major sources of the infections spread (Koo et al., 2020). According to an editorial published in the *New England Journal of Medicine (NEJM*), the asymptomatic spread of

the virus is the "Achilles' heel of enforced strategies to control COVID-19" (Gandhi, Yokoe, & Havlir, 2020). The basic SIR and SEIR models overpredict because they do not take into account the hospitalization and quarantine condition which severely reduces the overall spread of the infection. Thus, we have used the SEIQRD model to consider the hospitalization effect and capture disease transmission by the asymptomatic and pre-symptomatic cases.

Fig. 6.1 presents the SEIQRD epidemic model where S, E, I, Q, R, and D represent the susceptible (S), exposed (E: infected but not yet infectious), infectious (I), quarantined or hospitalized (Q), recovered (R), and deceased (D). An additional compartment I_2 (with retention period, T_2= 1 day) has been added to represent the newly infected cases on a particular day. The host total population is $N(t) = S(t)+E(t)+I_1(t)+ I_2(t)+Q(t)+R(t)+D(t)$ at time t. T_0 represents the incubation period, the duration between the viral exposures, and becoming infectious. The average incubation period reported from distinct sources is 5-6 days with a probable range of 2-14 days (WHO, 2020; Liang, 2020). T_1 denotes the infection period where a person is infectious but not symptomatic, whereas, T_3 indicates the recovery period.

The model is based on two assumptions: (i) Natural births as well as deaths due to other reasons during the study period are not considered in counting the total population, (ii) the recovered people are immune to further viral attacks during the short span of analysis. The second assumption is clinically proven by many studies for various contagious viral attacks e.g., (Zhou, Miranda-Saksena, & Saksena, 2013; Short, Kedzierska, & Van de Sandt, 2018). Eqs. 6.1-6.7 present the mathematical formulations of the SEIQRD epidemic model. The symbol φ_1 represents the fraction of symptomatic infections whereas φ_2 denotes the mortality fraction of the quarantined/hospitalized cases. The contagion rate (a), infection rate (c), and recovery rate (e) have been presented in Eq. 6.8-6.10.
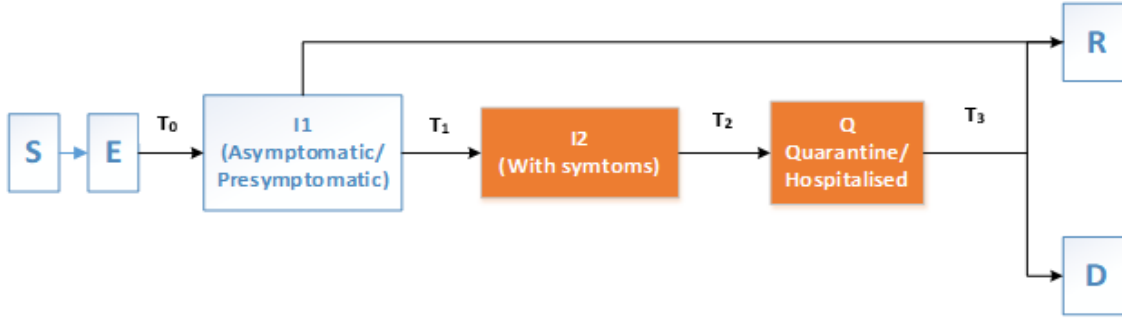
Fig. 6.1: Schematic representation of the SEIQRD model

$$\frac{dS(t)}{dt} = -\frac{aS(t)I(t)}{N} \tag{6.1}$$

$$\frac{dE(t)}{dt} = \frac{aS(t)I(t)}{N} - bE(t) \tag{6.2}$$

$$\frac{dI_1(t)}{dt} = bE(t) - cI_1(t) \tag{6.3}$$

$$\frac{dI_2(t)}{dt} = \varphi_1 cI_1(t) - dI_2(t) \tag{6.4}$$

$$\frac{dQ(t)}{dt} = dI_2(t) - eQ(t) \tag{6.5}$$

$$\frac{dR(t)}{dt} = (1 - \varphi_1)cI_1(t) + (1 - \varphi_2)eQ(t) \tag{6.6}$$

$$\frac{dD(t)}{dt} = \varphi_2 eQ(t) \tag{6.7}$$

Rate of contagion, $a \left(\frac{infection}{day}\right) = Number\ of\ contacts\ by\ infected\ person\ per\ day *$

$$each\ contact\ turning\ into\ infection \tag{6.8}$$

$$infection\ rate, c = \frac{1}{Infection\ period} = \frac{1}{T_1} \tag{6.9}$$

$$recovery\ rate, e = \frac{1}{Recovery\ period} = \frac{1}{T_3} \tag{6.10}$$

In terms of the basic reproduction number, the rate of contagion a could be presented by Eq. 6.11. The effective rate of contagion (a_{eff}) with the non-pharmaceutical interventions (NPI) could be given as Eq. 6.12,

$$a = \frac{R_0}{T_1} \qquad (6.11)$$

$$a_{eff} = \rho \frac{R_0}{T_1} \qquad (6.12)$$

where $'\rho'$ is the exposure factor. Its value changes with government regulations and individual practices.

The present model does not explicitly consider self-isolation of asymptomatic cases, contact tracing, and super-spreading events. These factors have been included implicitly using a lumped parameter approach. Also, the present work does not consider the distinct recovery time for mild cases and critical cases requiring ventilators and intensive care.

6.2.2 Evaluation of the parameters of the SEIQRD model

Fig. 6.2 presents approaches to solving the SEIQRD epidemic model by parameter tweaking and parameter learning using the artificial neural network (ANN). TT_0 is the time of the first death reported due to SARS-CoV-2. The data from TT_0 to TT_1 is used to determine the intermediate parameters by tweaking. We have taken 15 days between TT_0 to TT_1 assuming that the dynamics would be established in those 15 days and the parameters such as the incubation rate (b), the infection rate (c), and the recovery rate (e) would almost settle during this period. The alterations in the contagion rate (a) due to the miscellaneous public safety regulations and the individual practices would be supervised between TT_1 to TT_2. The minor changes in other parameters would also be recorded. The manual parameter tweaking provides a better fit when done for one set of data applying a trial and error approach. This is time-consuming and is not feasible for regular model updating as new data becomes available. The ANN recursively calibrates based on the data.

The ANN-based model calibration is a faster way of updating the model and can be used to calibrate the model on-demand as the new data becomes available.

The schematic diagram of the parameter fitting has been illustrated in Fig. 6.3. The parameters were initialized by the generic values given by the WHO-China joint mission report on COVID-19 (Aylward, Bruce, 2020; Liang, 2020). The cost function of the minimization for determining the model parameters has been presented in Eq. 6.13. The mortality data is a more reliable measure for predicting the severity of an epidemic. However, the models based on the initial mortality data cannot estimate surplus death caused due to the unavailability of sophisticated treatment if the infection peak exceeds the healthcare capacity. The infection cases are subjected to under-reporting due to numerous reasons e.g., inferior surveillance and tracking system, poor testing policy, asymptomatic cases, and distinct immune systems. We have employed a weighted cost function based on the mortality data and newly infected cases to obtaining robust parameters. A higher weight (w_2= 0.6-0.8) was assigned to the fatality data than the infection data (w_1= 0.2-0.4) assuming that the mortality data is more accurate.

$$f = \sqrt{w_1 * \left(I_{reported} - I_{Model}\right)^2 + w_2 * \left(D_{reported} - D_{Model}\right)^2} \qquad (6.13)$$
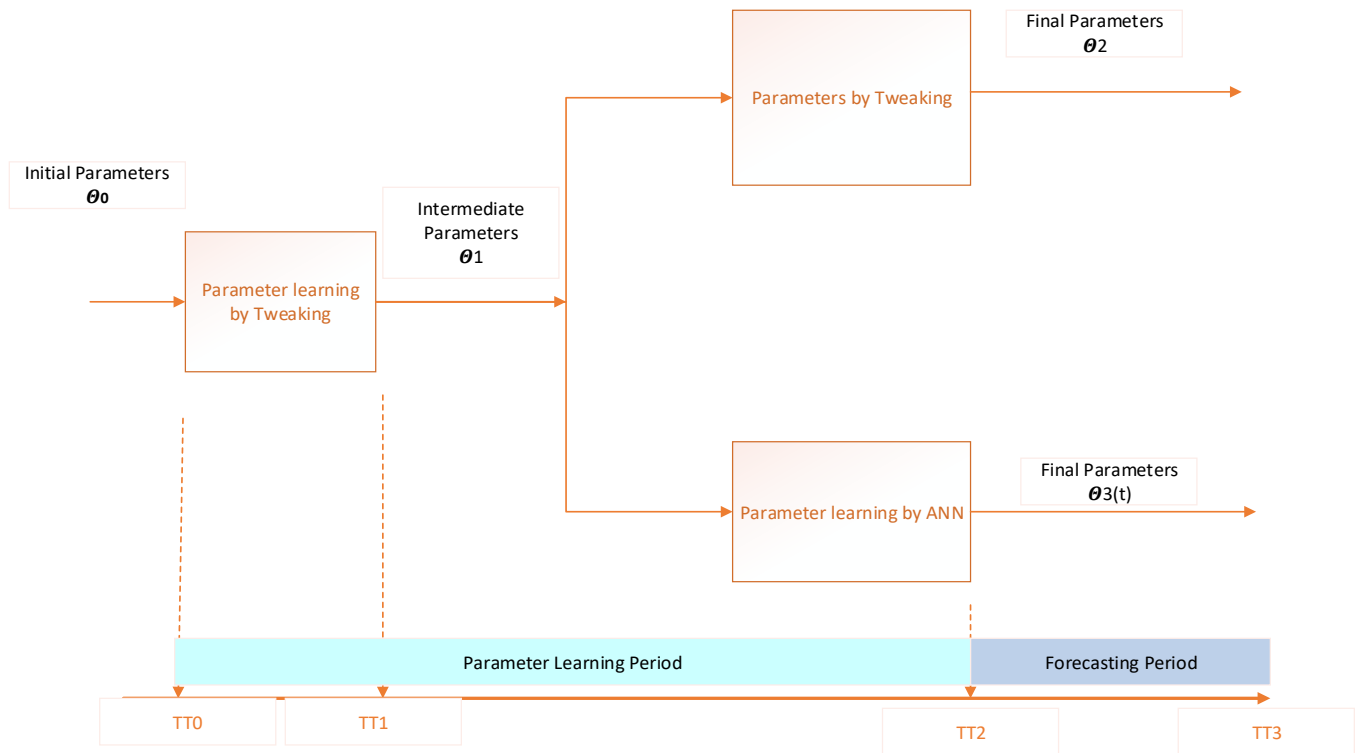
Fig. 6.2: Parameter learning of the SEIQRD model using parameter tweaking and ANN-based calibration
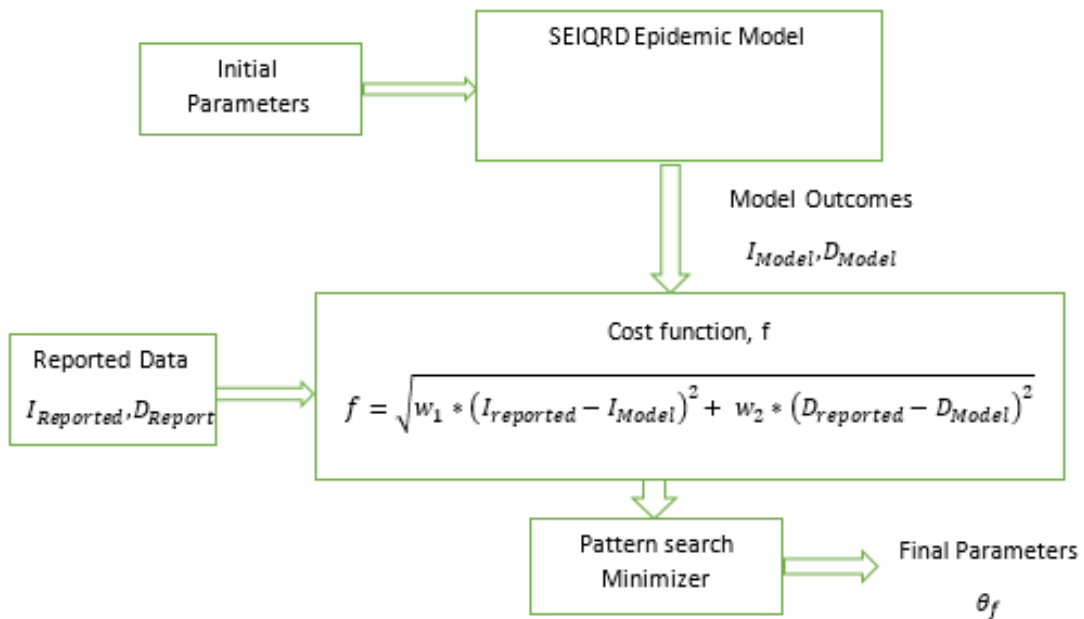
Fig. 6.3: Schematic representation of parameter fitting of the SEIQRD model

169

6.3 Results and discussion

6.3.1 The case study:

The proposed models have been used to study the present pandemic caused by severe acute respiratory syndrome coronavirus (SARS-CoV-2) that was first identified in December 2019 in Wuhan, Hubei province of China. The outbreak rapidly spread worldwide and was declared a global pandemic on March 11, 2020 (Kopecki et al., 2020; WHO, 2020).

The proposed models have been evaluated on forecasting the COVID-19 pandemic spread at four locations: Ontario, British Columbia, Italy, and Germany. British Columbia was among the first provinces in Canada to be affected by COVID-19. The province demonstrated good governance in public health management resulting in a low number of per capita death. Ontario is the most populous province in Canada with 14.7 million people representing 38.8% of the country's population (Ministry of Finance, Government of Ontario, 2019). Also, it is home to the country's most populous city, Toronto, and the capital city, Ottawa. Italy was the first western country impacted by the Coronavirus and experienced the longest lockdown of Europe. The robust response of Germany resulted in a vastly lower death rate despite the higher number of infections.

The models have been calibrated using the pandemic data reported by John Hopkins University (Johns Hopkins University Center for Systems Science and Engineering, 2020). Table 6.2 lists the generic parameters of the model. The average incubation, infection, and recovery periods were 5.5, 5.1, and 11.5 days, respectively. The basic reproduction number for the COVID-19 reported by multiple sources has been presented in Appendix A. The basic reproduction number of 2.9, the median values of all the sources, has been used as the initial guess for the tweaking model.

Table 6.2: Generic values of the model parameters

Parameter	Value (in days)	Parameter	Value
Incubation period	5.1	Recovery period (in days)	11.5
Infection period	5.5	Basic reproduction number	2.9

6.3.2: Model application in the forecasting of infections

Fig. 4 presents cumulative infections caused by COVID-19 at selected regions. The models were trained up to T=40 days and used for forecasting for extended periods. The accuracy of distinct models is measured in terms of the root mean squared error (RMSE). The RMSE is a measure of the difference between the calculated and the observed values and is computed as $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(I_{reported,i} - I_{Model,i})^2}$. Fig. 6.4a depicts the forecast for Ontario by the generic model, the parameter tweak model, and the learning-based model. The RMSE values for the generic model, parameter tweak model, and the continuous learning ANN-based model were found to be 9.18 x 10^3, 8.87 x 10^3, and 3.15 x 10^3 respectively. It is observed that the generic model overpredicts the infected cases. A plausible reason for lower real infections could be due to social distancing and other intervention strategies already in effect during the period of study. The government enforced preventive measures such as limiting gathering sizes, closure of non-essential businesses and schools, and emergency lockdown. The population is following public advisory such as maintaining social distancing, frequent hand washing and practicing hygiene. As a result of all these, the true cases went down. The other factors can be a lack of adequate testing and under-reporting of the infected cases. This is captured by the tweaking and learning-based models. The parameter tweaked model registers better performance compared to the generic

model. The continuously learning ANN-based model yields the lowest RMSE among the test methods. This is due to the dynamic learning of the parameters in varying conditions.

Fig. 6.4b and Fig. 6.4c present the comparative performance of the tweaking and the learning-based models. The tweaking model functions well in the training range, however, it shows poor performance in forecasting for an extended period. The learning-based methods precisely capture the dynamic variations in the parameters resulting in realistic projections. Similarly, the ANN based-learning model is effectively forecasting the cumulative infections of British Columbia (Fig. 6.4d), Italy (Fig. 6.4e), and Germany (Fig. 6.4f). It is plausible to realize a deviation of the model outcomes from the observed data if some other regulations were not there during the training period. The performance will also be affected by the variation of individual or societal practices that were not well established during the training phase. It can be noticed that the model is slightly overpredicting for the cases of Germany. This might be due to their COVID-19 monitoring initiatives, Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) and the Corona Data Donation app (Corona-Datenspende).

The models were also employed to assess newly infected cases of COVID-19 considering randomness in the model parameters. We employed Monte Carlo simulations based on the lognormal distributions of the incubation, infection, and recovery periods to capture the long tail of the infection risk (Fig. 6.5). The distributions of the incubation period, the infection period, and the recovery period are presented in Fig. 6.5a, Fig. 6.5b, and Fig. 6.5c, respectively. The infection per day (the new infections) grows moderately at the beginning, attain maximum, and then start descending. The peak of the newly infected cases is crucial in developing treatment strategies for infected populations. It should not surpass the existing healthcare facility of the corresponding territory. A precise estimate of the peak infection in advance is conducive to better preparation and

effective mitigation. Distinct measures should be adopted to lower the peak, also known as flattening the epidemic curve, for restraining the pandemic. Fig. 6.5d presents the forecast for the peak of infection for Ontario. It can be inferred that the average value of the peak infection is around 1000 cases per day. The peaks of the infections were varying between 225 and 2500 within 25% and 75% quantiles. Similarly, the most probable value of the cumulative infection is 4.0×10^4. The maximum cumulative infection could go up to 8.0E+04 considering uncertainty in the parameters.
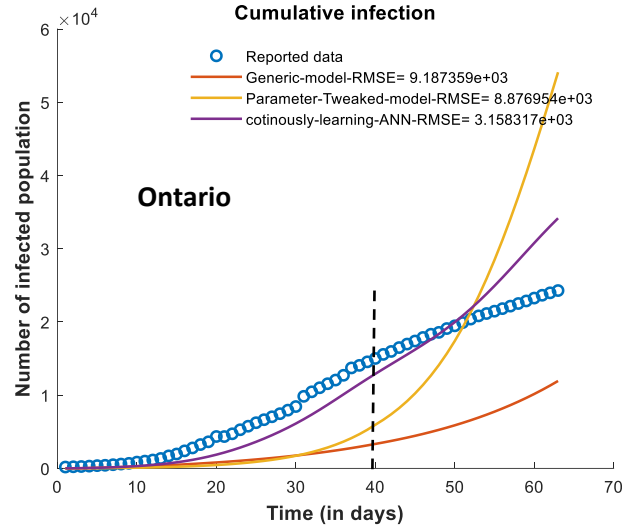
a. Cumulative infection forecast of distinct methods
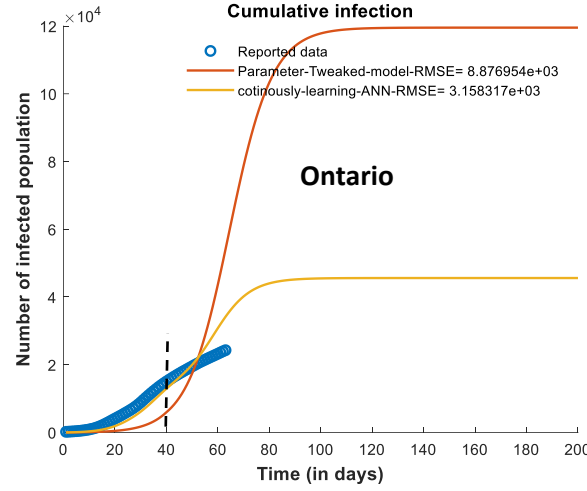
b. forecast using tweaking and ANN-based learning methods
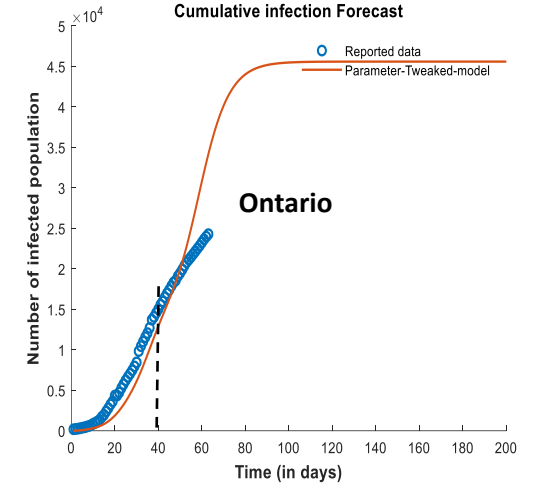
c. forecast using ANN-based learning model

d. Cumulative infection forecast of the ANN-based learning model

e. Cumulative infection forecast of the ANN-based learning model

f. Cumulative infection forecast of the ANN-based learning model
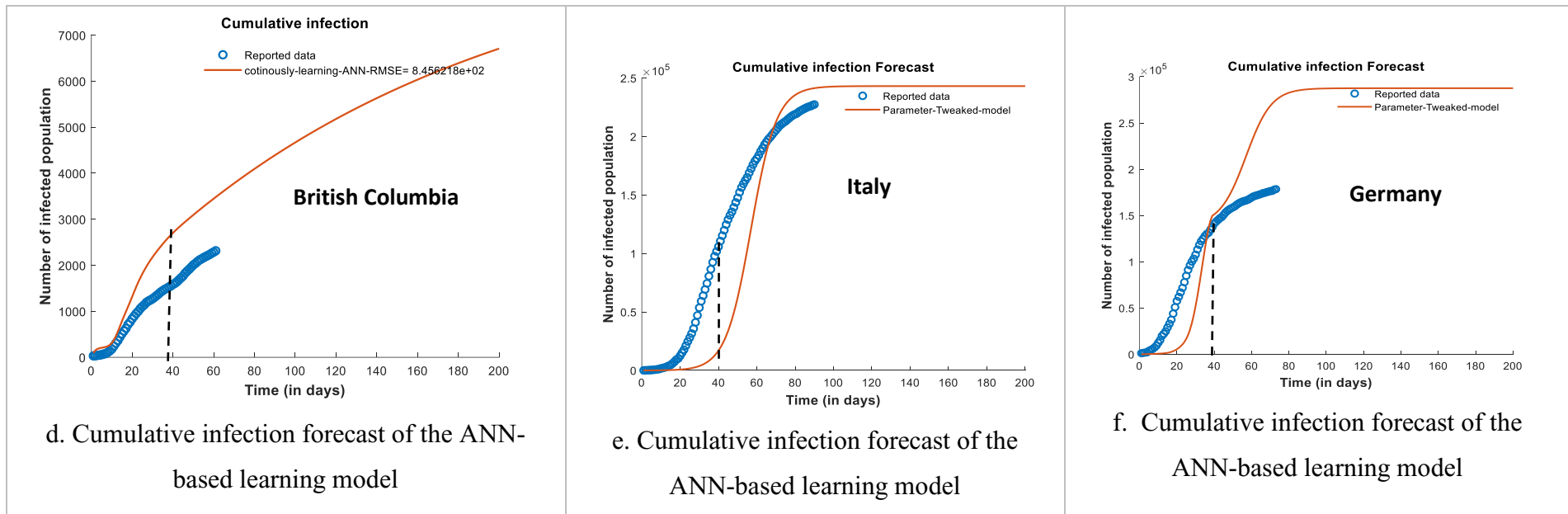
Fig. 6.4: Forecast of the infected population of COVID-19 at selected regions using the SEIQRD model (The dash line at T=40 represent the training period of the models for the forecast of the extended period)
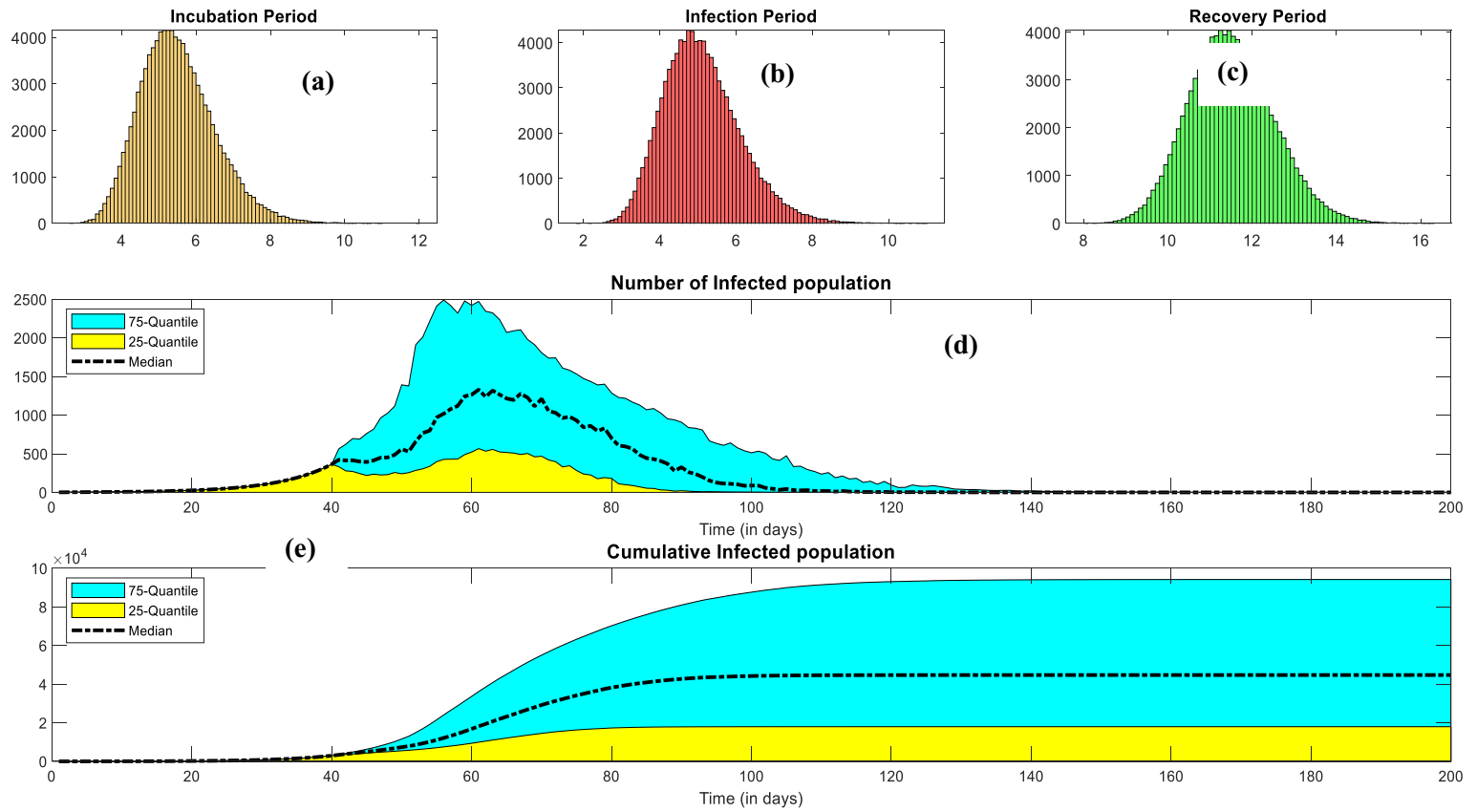
Fig. 6.5: Predicting infection risk considering randomness in the incubation, infection, and recovery periods

(a) Incubation period, (b) infection period, (c) recovery period, (d) peak infection per day, (e) cumulative infection

6.3.3 Model application in managing risk of an epidemic

The transmissibility of the epidemic is quantified by basic reproduction number (R_0) which is defined as "the average number of secondary cases generated by a primary case in an entirely susceptible population" (Ferguson et al., 2005). The epidemic spreads for $R_0 > 1$ and dies out if $R_0 < 1$. The effective reproduction number can be reduced in three ways: i) reducing contact rates, ii) diminishing the infectiousness through isolation and treatment, and iii) reducing the susceptible population by vaccination (Ferguson et al., 2005).

Personal hygiene such as wearing a mask at public places, frequently washing hands and social distancing are crucial in restraining the spread of the infection risk. Government regulations and individual practices play a crucial role in mitigating epidemic spread. Limiting gathering sizes, closure of nonessential businesses and schools, and emergency lockdown have a decisive role in controlling the epidemic spread (Davies et al., 2020; Ferguson et al., 2006). Restricting limits on gathering sizes reduces the possibility of super-spreading events (Anderson, Heesterbeek, Klinkenberg & Hollingsworth, 2020). Following social distancing measures prevent transmission by pre-symptomatic and asymptomatic cases. Vaccination is compelling in protecting from infectious diseases; however, several months are required to develop, produce, and distribute an effective vaccine following the outbreak of a novel pandemic strain (Germann et al., 2019). Thus, NPIs are vital in reducing overall mortality by restricting peak within the healthcare capacity.

The predicted effects of interventions on the spread of the epidemic have been presented in Fig 6.6. Fig. 6.6a shows the impact of the pandemic if no measure is taken to restrict the outbreak. Fig. 6b and Fig. 6c respectively demonstrate the effect of the school and non-essential business closure and emergency lockdown after one week of the first mortality reported. Similarly, the effects of these interventions if implemented after one month of the first mortality, have been

presented in Fig. 6.6d and Fig. 6.6e. The Monte Carlo simulation has been used for considering uncertainties of the model parameters. We can observe that the median value of the peak infection could reach up to 3000 cases per day in case of no intervention. It could go up to 10000 cases per day in the worst-case scenario. The closure of schools and businesses and observing public emergency reduces the new cases to 6 per day as the most probable value. The lockdown results in rapid shrinkage of the epidemic compared to simply school and nonessential business closure. We can also deduce that delaying interventions by a month can lead to a several-fold increase in the peak of the infection.

The effect of the pandemic in terms of peak risk has been shown in Table 6.3. It presents the risk estimated for different locations in terms of risk peak and the cumulative risk of infections. The risk can be defined as the product of the impact of the pandemic and the probability of the occurrence of the impact. In this SARS-CoV-2 pandemic, the impact is infections which could lead to mortality. We are studying the most probable value of infections as the potential impact of the COVID-19. Monte Carlo simulation has been used for handling uncertainties in the number of infection cases due to the pandemic. The probability of occurrence of infection has been computed from the distribution of infections considering the randomness of the model parameters (i.e., the incubation, infection, and recovery periods). The probability of the most probable value denotes the possibility of realization of the most probable number of infections for a given time. It has been calculated based on the distribution of impacts. For instance, the value of the most probable impact at T=50 was found to be 2867 cases per day. The frequency of the most probable impact for T=50 is 433 out of 1000 possible values resulting in the occurrence probability of the most probable impact of 0.433 ($\frac{433}{1000} = 0.433$). Thus the risk of infection at T=50 will be 2867*0.433=1241 cases per day. The values of parameters of Fig. 6.7a-b at other time steps have been computed similarly.

In the case of no measures to restrain the COVID-19, the respective peak risk (infections per day) and the cumulative risk (total infections) in Ontario are 1.39×10^3 and 4.26×10^4. The risk will be reduced remarkably by enforcing non-pharmaceutical interventions. Lockdown is effective in quickly containing a pandemic. Also, a delay in implementing intervention strategies increases the risk markedly. Table 6.3 shows the comparative outcomes of intervention strategies implemented after a week and a month from the first death reported. The peak risk increases from 3 to 43 (infections per day) while the cumulative risk escalates from 122 to 2.07×10^3 (total infections) in the case of a delay by a month in administering school/university and non-essential business closures. Similarly, an increase from 3 to 39 (infections per day) in peak risk and 22 to 5.64×10^2 (total infections) in cumulative risk is observed by a month-long delay in enforcing the lockdown. The marked escalation in the peak and the cumulative risk of the other locations could also be seen due to the delay in implementing the intervention strategies.
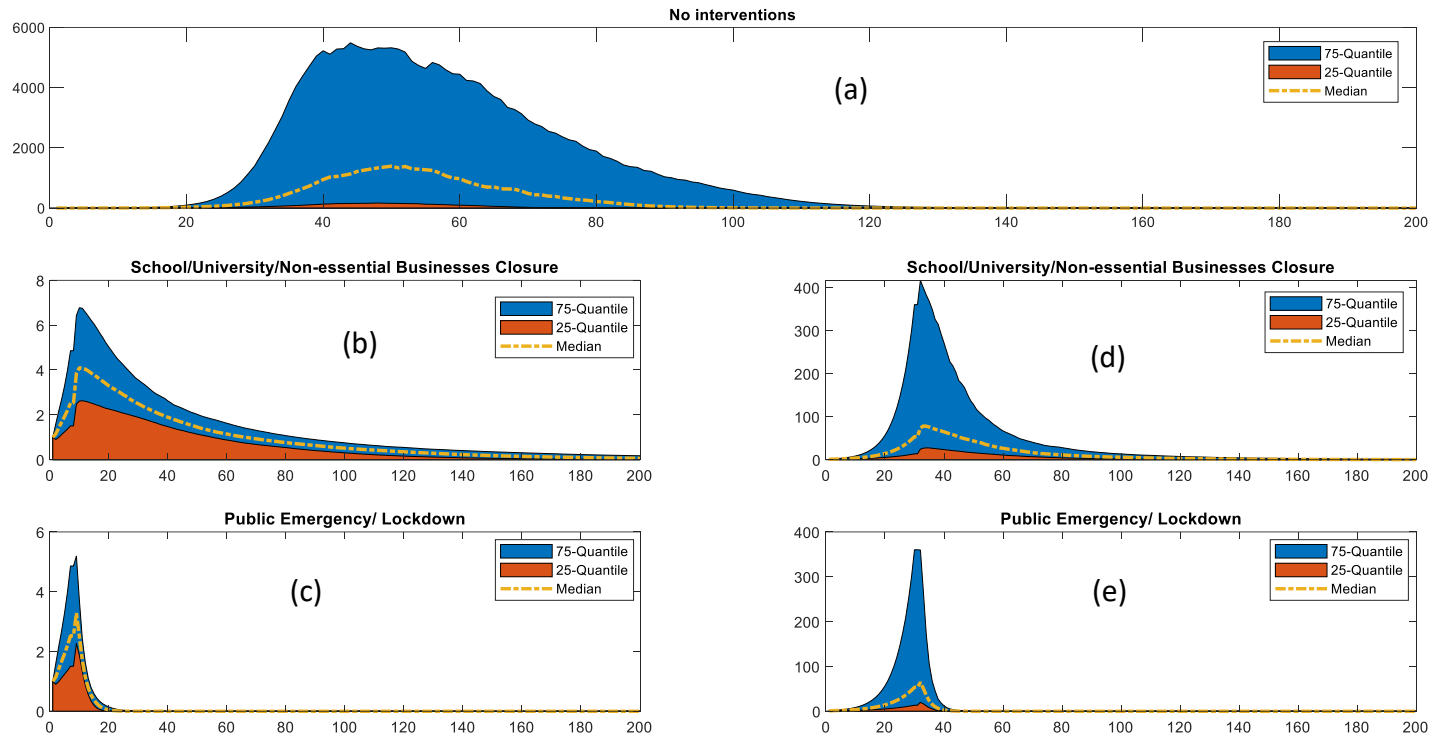
Fig. 6.6: Effect of interventions on controlling the epidemic risk; a: Without any intervention, b: School and non-essential business closure after one week of the first mortality), c: Enforcing public emergency/ lockdown after one week of the first mortality d: School and non-essential business closure after one month of the first mortality), e: Enforcing public emergency/ lockdown after one month of the first mortality

Table 6.3: The risk of epidemic infections in selected regions

(Risk of infection= Impact in terms of infections x probability of occurrence of the impact)

Countries/ cities	Parameters	No Measure	Interventions implemented after one week		Intervention implemented after one month	
			School/ Universities/ non-essential business closure	With Public Emergency/ Lockdown	School/ Universities/ non-essential business closure	With Public Emergency/ Lockdown
Ontario	Peak Risk (Infections per day)	1.39×10^3	3.00×10^0	3.00×10^0	4.30×10^1	3.90×10^1
	Cumulative Risk (Total Infections)	4.26×10^4	1.22×10^2	2.00×10^1	2.07×10^3	5.64×10^2
British Columbia	Peak Risk (Infections per day)	4.47×10^3	48.00×10^0	48.00×10^0	9.99×10^2	6.53×10^2
	Cumulative Risk (Total Infections)	1.29×10^5	1.70×10^3	3.28×10^2	1.53×10^4	6.45×10^3
Italy	Peak Risk (Infections per day)	8.62×10^5	9.53×10^4	4.00×10^0	4.55×10^5	1.20×10^1

	Cumulative Risk (Total Infections)	1.18×10^7	3.66×10^6	7.60×10^1	8.08×10^6	2.70×10^2
Germany	Peak Risk (Infections per day)	4.26×10^4	3.00×10^0	3.00×10^0	4.26×10^4	4.26×10^4
	Cumulative Risk (Total Infections)	2.31×10^5	1.40×10^2	2.20×10^1	2.31×10^5	2.31×10^5
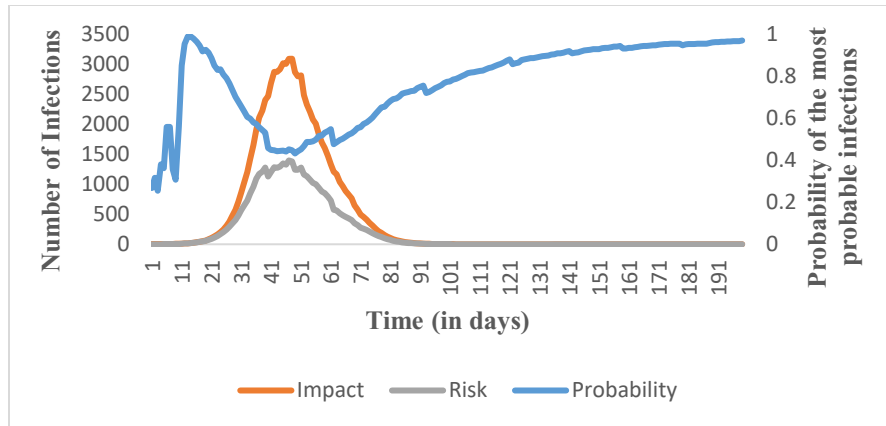
Fig. 6.7a: Variation of peak values of the number of infections in terms of the most probable impact and the calculated risk
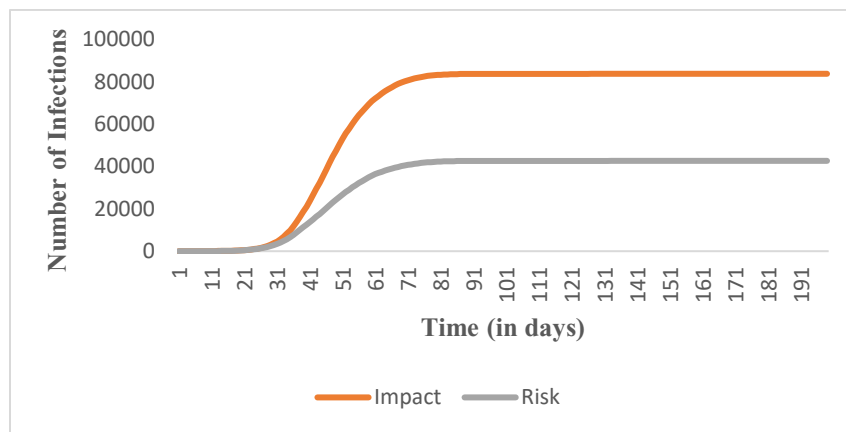
Fig. 6.7b: Variation of cumulative values of the number of infections in terms of most probable impact and the calculated risk

6.3.4 Model application in evaluating risk in releasing scenarios

The lockdown and other government regulations have severe economic impacts. It is not viable to continue with these regulations for a prolonged duration. Thus, the enforcement and release of such regulations must be exercised judiciously to prevent harmful impacts of a pandemic. Fig. 6.8 presents projection of releasing the enforced interventions. The relaxing scenarios of opening school/university/non-essential business, and relaxing on social gatherings for returning to the normal state have been modeled for forecasting its impact on the overall system. The relaxing regulations on social gatherings comprise the opening of the parks, malls, bars, religious

183

gatherings, recreation facilities, and all the sites of the interactions other than working sites. Fig. 6.8 depicts the comparative escalation in terms of the median values of the peak infections. It also presents the 25 and 75 quantile of the peak infections to capture the alteration due to the randomness in the model parameters. It is inferred that the release of the enforced regulations has the potential for a resurgence. The phase-wise relaxation of these regulations is most appropriate for preventing the resurgence of the pandemic risk. Relaxing regulations on social gathering at T=70 could lead to a resurgence with the peak infection capacity of twice that in the regulated scenario. However, the opening of schools/nonessential business and relaxing in social gatherings at T=70 and T=100, respectively, do not exhibit significant resurgence.
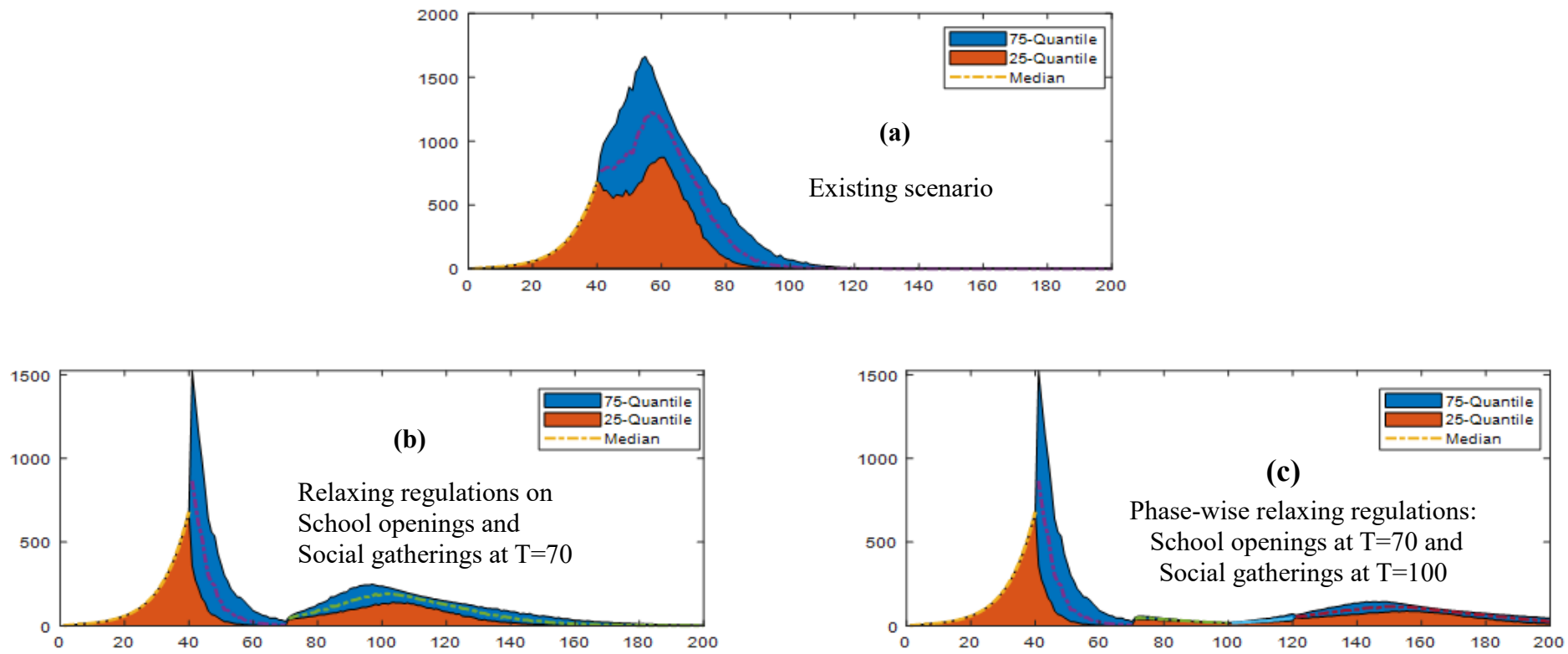
Fig. 6.8: Effect of relaxing regulations on the impact of the pandemic

6.3.5 Similarity between pandemic risk management and process safety management

The current pandemic situation and its prevention strategy can interchangeably be used for hazard identification and safety management of processing facilities. Forecasting of the infection strategy stated above is applicable to identify a possible abnormal scenario. The risk-based safety management framework shown above illustrates the importance of safety management layers when an abnormal situation is triggered. The distinct measures of handling a pandemic can be represented as the layer of protection analysis of a process system (Fig 6.9). The processes are designed safer as a preventive strategy for reducing the risk. Different safety barriers such as basic process control, alarms and operator interventions, safety instrumented system, relief devices, and physical containments are used as a control layer for abnormal situation management (Dowell, 1999; Willey, 2014). Finally, the plant and emergency response services are brought into operation to diminish risk by de-escalating the situation. These are analogous to the preventive, controlling, and mitigative strategies used for handling a pandemic. The hygienic practices (e.g. frequent hand washing, wearing a mask at public places) and government regulations such as the closure of schools and non-essential business, limiting gathering sizes, enforcing lockdown, and vaccination is extremely effective as control and preventive strategies for limiting pandemic risk. Plant and community emergency responses are mitigative layers for industrial systems. These control measures are analogous to the treatment and medical care provided for reducing the impact of the pandemic.

Results in Fig 6.6b-e show that effective control measures can reduce the risk of the disease spread. The results also demonstrate how early intervention of the preventive measures can reduce risk. Thus, protection layers should be put into place as soon as a hazard is identified.

We have also illustrated the impact of relaxing scenarios in Fig. 6.8b-d. Releasing different barriers causes the risk to rise. The time of the release of the barrier plays a crucial role. The risk reduces significantly when a barrier is released after the peak of the occurrence. Similarly, risk-based analysis of safety barriers could be feasible in process systems to determine how long a critical control layer should be enforced, and when it needs to be released.
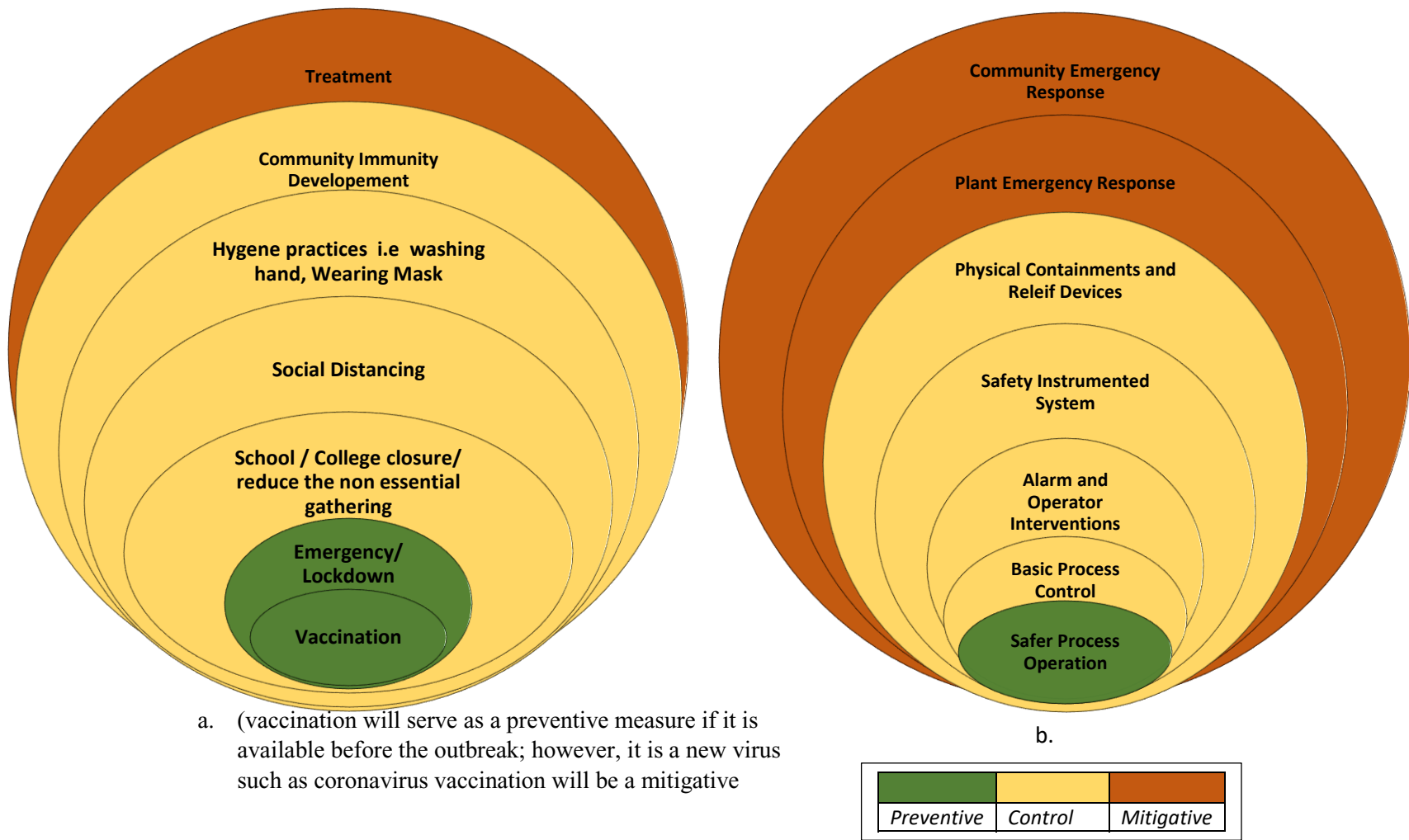
Fig. 6.9: Layer of protection analysis (LOPA) for epidemic and abnormal situation management in a process system; a. for epidemic management; b. for the safety of a process system

The COVID-19 pandemic is a disastrous event that is affecting billions of lives and causing adverse economic impacts. World Bank economic reports indicate that a severe pandemic could reduce world gross domestic product (GDP) drastically (World Bank, 2020). Nonetheless, it provided many learning opportunities to policymakers and process safety practitioners. Some of those lessons are as follows-

Early fault detection vs early case detection

The early warning system for hazard identification is central to abnormal situation management. A delay in detection would lead to a delay in control actions that will escalate risk resulting from abnormal situations. For instance, a delay of just a few days in releasing genetic sequences can be critical in an outbreak. Experts believe that the risk would have been reduced dramatically if the outbreak was detected antecedently. This is equally applicable to a process system where early detection of a fault reduces the risk of potential loss.

Identification of effective control mechanism

The identification of distinct control strategies is conducive to restrain abnormal situations. The restrictive measures e.g., social distancing, case detection, isolation, contact tracing, quarantine of exposed cases, and lockdown turns out to be effective strategies for restricting the spread of the COVID-19. Good hygiene practices such as frequent hand washing, wearing a mask at public places, and physical distancing are also effective in flattening the curve and reducing the economic burden. Thus identification of an effective control mechanism plays a pivotal role in minimizing the impact of an abnormal situation.

The fast response of public health vs operator response

We can also deduce that the delay in the intervention strategies significantly increases the risk. We have analyzed the catastrophic effect of the delay in the implementation of selected intervention strategies in our model. We found severalfold escalation in peak and the cumulative risk of infection due to a delay of a month in implementing intervention strategies. The real-world data from the COVID-19 also substantiate our findings. For instance, the immediate response by the government of Taiwan, for handling people arriving from Wuhan was instrumental to reduced risk in Taiwan. The findings uphold the importance of the operator's response to an abnormal situation. As a delay to respond to a fault can cause detrimental consequences, prompt action from the operator is vital to ensure the safety of processing facilities.

Effective resource allocation and mobilization

Resource allocation decisions were critical segments of handling the COVID-19. The mobilization of resources in administering social distancing by school closures, limiting gathering sizes, providing efficient quarantine centers was of utmost importance in fighting the present pandemic. We had to properly allocate ventilators or intensive care beds in case of limited availability. We also experienced that many doctors, nurses, and health workers fell victim to the pandemic. Thus, we can pledge for the exclusive treatment for healthcare workers, technicians, and security personnel who risk their lives as front liners (Khoo & Lantos, 2020). Equivalently, we can identify and provide maintenance to the pivotal elements of process safety systems.

Adoption of advanced technologies

The use of advanced technology was helpful at all levels of containment. South Korea credited its use of advanced technology to uncover clusters of cases that would otherwise have gone undetected. Drones equipped with cameras hovering over parts of the Indian neighborhoods

warning residents they are under surveillance turned out to be advantageous in enforcing effective lockdown. Artificial intelligence is being tested for identifying the disease by recognizing patterns in ultrasonic images. The AI can also be useful in understanding the virus and accelerating medical research on vaccine development and treatments. The advanced data mining tools can uncover the virus's history, transmission, diagnostics, management measures, and extracting features for combating future epidemics. Virtual assistants and chatbots have been deployed to support healthcare systems in many countries such as Canada, France, Finland, Italy, Germany, and the United States. Thus, advanced technologies can be harnessed for reducing risk and abnormal situation management in process systems.

Identification of vulnerable elements

The COVID-19 also revealed that some people are most likely to be the hardest hit by the current pandemic. The most vulnerable and high-risk group include: an older adult suffering from underlying medical conditions (e.g. heart disease, hypertension, diabetes, chronic respiratory diseases, cancer) and individuals having a compromised immune system from a medical condition or treatment (e.g. chemotherapy). We provided extra care for the vulnerable population for reducing the risk. We can identify and develop effective strategies and action plans for vulnerable elements for reducing risk in process systems.

New opportunities, scale-up, and resiliency of the existing systems

The COVID-19 highlighted scaling-up requirements of health systems to expand services to accommodate rapid increases in demand. We also experienced numerous innovative initiatives during the COVID-19 pandemic. Many companies that don't regularly operate the business of medical products, started producing hand sanitizer, ventilators, or personnel protection equipment (PPEs). Many businesses, schools, universities, and other organizations demonstrated resiliency

by continuing their operations by quickly adopting the new normal. They widen the virtual private network (VPN) to allow all employees to work remotely; brought changes in the existing system by the necessary modification such as adding physical barriers at the help desk and cash counters; moving to online portals for meetings and instructions.

Good governance and trust-building

The COVID-19 made us realize the importance of trust-building and effective governance in fighting a catastrophe. The distinct government regulations are prone to create clashes between states and citizens, eroding state capacity, driving population displacement, and heightening social tension and discrimination. The clashes can potentially escalate an abnormal situation leading to a catastrophe. The safe operation of process systems depends on the smooth functioning of equipment, operators, and managerial systems. An effective management system can improve health and safety at work by optimizing human interactions in the technical and social environment through proper policies, procedures, training, and supervision.

Application of expertise from similar outbreaks in past vs use of historical process data

Operators use their experience to detect, diagnose, and take mitigating action for controlling abnormal situations in process systems. In addition to this, the data-based fault detection and diagnosis methods are trained based on historical faulty data and expert opinions. Despite the different outcomes of epidemic disease, the expertise from past outbreaks could be conducive to a credible estimation of the trajectory and slowing down the spread by implementing effective measures (Goudarzi, 2020, March 23). The social distancing measures, communication, and international cooperation, the most effective methods to slow COVID-19, were adopted by experience from the 1918 influenza pandemic and 2002–2003 SARS outbreak. The expertise from the past outbreaks e.g., the 2003 severe acute respiratory syndrome (SARS) outbreak in Singapore,

and experience with the 2015 Middle East respiratory syndrome (MERS) outbreak of South Korea led to an immediate fruitful response to the COVID-19. Their approach in combating COVID-19 was praised as "the gold standard" response (Firth, 2020, March 05).

Pandemics do not remain geographically confined in contrast to other natural disasters (Jamison et al., 2017). A holistic approach with strong ethical and sensible measures is required for combating the epidemic spread (Institute of Medicine, 2007). We have to be prompt in all facets of the transmission; adequate testing facilities, active surveillance, enforcing intervention strategies, and community screening around the cluster areas. The extensive support and public endorsement can be asserted by effectively communicating the preparedness and response strategies. Research findings should be disseminated in the form of actionable points such as checklists **(**Khalid, 2020, March 03**).** The migration and other cross-border entries pose the risk of further spreading an outbreak; it must be handled effectively. All the aforementioned ideas could potentially be employed for upgrading the safety of process systems.

6.4 Conclusions

In the current work, a data-driven semi-mechanistic SEIQRD model was used to develop a risk management framework to forecast the spread of the COVID-19 pandemic. A parameter tweaking and ANN-based learning were employed to evaluate the adaptive parameters of the model. The proposed models were able to forecast the spectrum of disease transmission at an early stage. Pandemic data from four different geographical locations were used to demonstrate the efficacy of the presented frameworks. The proposed model predicted cumulative infections with different safety barrier implementation and release scenarios. The risk incurred was evaluated on implementing and relaxing different safety barriers. The results demonstrated that implementing

non-pharmaceutical interventions can dramatically reduce the risk. The time of the enforcement and relaxing of NPIs also play a crucial role in restraining the epidemic risk. The NPIs can turn ineffective if the implementation is delayed or relaxed abruptly.

The proposed risk-based method has many similarities with process facilities for hazard identification and safety management. In the current work, the implementation of the LOPA framework for managing public health safety and risk was studied. A risk-based analysis was performed for three different scenarios: (i) no protection layer added (ii) early enforcement of protection layer (iii) delayed enforcement of protection layer. Early enforcement of a protection layer is crucial to keep the risk significantly lower. The proposed study suggests that the enforcement and relaxing of the protection layer should precisely be executed based on reliable forecasting of the model.

The proposed model can perform well when calibrated for specific regions using local data and information such as population, demographics, interaction patterns, enforced regulations, and other dynamics. The present model does not explicitly consider contact tracing and super-spreading events. These factors have been included implicitly using a lumped parameter approach. Also, the present work does not consider the distinct recovery time for mild cases and critical cases requiring ventilators and intensive care. The model could also be improved by dividing populations based on demographics, special dispersion, and interaction patterns. Effects of distinct degrees of social distancing, wearing a mask at public places, and following hygiene practices can also be analyzed. Finally, the model could be used in trade-off studies for balancing economic aspects and acceptable risk when enforcing and relaxing regulations.

Acknowledgment

List of symbols and abbreviations

Symbols	Meanings
A	contagion rate
ANN	artificial neural network
C	infection rate
C_A, C_B	concentrations of species A and B respectively
COVID-19	Coronavirus disease of 2019
CSTR	continuous stirred tank reactor
D, D(t)	Decesead
D_{Model}	estimated fatality bt the model
$D_{Reported}$	reported death
E	recovery rate
E, E(t)	Exposed
I, I(t)	Infected
I_{Model}	infected cases estimated by model
$I_{Reported}$	reported infected cases
k_1, k_2	rate constants
LOPA	layer of protection analysis
M	total Mass
MERS	Middle East respiratory syndrome
N, N(t)	total population
NPIs	non-pharmaceutical interventions
PEPP-PT	Pan-European Privacy-Preserving Proximity Tracing
Q, Q(t)	Quarantined
R, R(t)	Recovered
R0	basic reproduction number
RMSE	root mean squared error
S, S(t)	Susceptible
SARS	severe acute respiratory syndrome
SARS-CoV-2	severe acute respiratory syndrome coronavirus-2

SEIQRD	susceptible, exposed, infected, quarantined, recovered, deceased
SEIR	susceptible, exposed, infected, and recovered
SIR	susceptible, infected, recovered
T1	infection period
T2	infection period
T3	recovery period
TT_0, TT_1, TT_2	timeline on the disease progression
w1, w2	relative weightage of the respective infected cases and fatality factor in determining the cost in the optimization
WHO	World Health Organization
a_{eff}	effective rate of contagion
φ_1	fraction of symptomatic infections
φ_2	fraction of quarantined/hospitalized population resulting in mortality
ρ	exposure factor
θ_0	initial parameters of the SEIQRD model
θ_1	intermediate parameters of the SEIQRD model
θ_2	final parameters of the SEIQRD model using tweaking
θ_3	final parameters of the SEIQRD model using ANN

References

Alvarado, A., Vedantam, S., Goethals, P., & Nopens, I. (2012). A compartmental model to describe hydraulics in a full-scale waste stabilization pond. *Water Research*, *46*(2), 521–530. https://doi.org/10.1016/j.watres.2011.11.038

Anderson, R. M., & May, R. M. (1979). Population biology of infectious diseases: Part I. *Nature, 280, 361-367*. https://doi.org/10.1038/280361a0

Anderson, R. M., Heesterbeek, H., Klinkenberg, D., & Hollingsworth, T. D. (2020). How will country-based mitigation measures influence the course of the COVID-19 epidemic? *The Lancet*, *395*(10228), 931–934. https://doi.org/10.1016/S0140-6736(20)30567-5

Aylward, Bruce (WHO); Liang, W. (PRC). (2020). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19). *The WHO-China Joint Mission on Coronavirus*

Disease 2019, *2019*(February), 16–24. Retrieved from https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf

Bermingham, S. K., Kramer, H. J. M., & van Rosmalen, G. M. (1998). Towards on-scale crystalliser design using compartmental models. *Computers & Chemical Engineering*, *22*(SUPPL.1), S355–S362. https://doi.org/10.1016/S0098-1354(98)00075-1

Busenberg, S., & Driessche, P. (1990). Analysis of a disease transmission model in a population with varying size. *Journal of Mathematical Biology, 28, 257-270*. https://doi.org/10.1007/BF00178776

Cui, Y. Q., Van Der Lans, R. G. J. M., Noorman, H. J., & Luyben, K. C. A. M. (1996). Compartment mixing model for stirred reactors with multiple impellers. *Chemical Engineering Research and Design*, *74*(2), 261–271.

Davies, N. G., Kucharski, A. J., Eggo, R. M., Gimma, A., Edmunds, E., J. (2020). The effect of non-pharmaceutical interventions on COVID-19 cases , deaths and demand for hospital services in the UK : a modelling study. The Lancet, 5(7), E375. https://doi.org/10.1016/S2468-2667(20)30133-X

Dowell, A. M. (1998). Layer of protection analysis for determining safety integrity level, ISA Transactions, *37(3)*, 155–165. https://doi.org/10.1016/S0019-0578(98)00018-4

Dowell, A. M. (1999). Layer of protection analysis and inherently safer processes. *Process Safety Progress*, *18*(4), 214–220. https://doi.org/10.1002/prs.680180409

Ferguson, N. M., Cummings, D. A. T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., Burke, D. S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia, *Nature, 437*(September), 209–214. https://doi.org/10.1038/nature04017

Ferguson, N. M., Cummings, D. A. T., Fraser, C., Cajka, J. C., Cooley, P. C., & Burke, D. S. (2006). Strategies for mitigating an influenza pandemic, *442*(July), 448-452. https://doi.org/10.1038/nature04795

Ferguson, N. M., Laydon, D., Nedjati-gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., Elsland, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P.G., Walters, C., Winskill, P., Whittaker, C., Donnelly, C., A., Riley, S., Ghani, A. C. (2020). Impact of non-pharmaceutical

interventions (NPIs) to reduce COVID- 19 mortality and healthcare demand, London: Imperial College London, March 16.

Firth, S, (2020, March 05). Singapore: The Model for COVID-19 Response? *MedPage Today* , https://www.medpagetoday.com/infectiousdisease/covid19/85254

Gandhi, M., Yokoe, D. S., & Havlir, D. V. (2020). Asymptomatic Transmission, the Achilles' Heel of Current Strategies to Control Covid-19. *New England Journal of Medicine,* 382:2158-2160. https://doi.org/10.1056/nejme2009758

Germann, T. C., Gao, H., Gambhir, M., Plummer, A., Biggerstaff, M., Reed, C., & Uzicanin, A. (2019). School dismissal as a pandemic influenza response : When, where and for how long ? *Epidemics*, *28*(June), 100348. https://doi.org/10.1016/j.epidem.2019.100348

Goudarzi, S, (2020, March 23). Lessons from Past Outbreaks Could Help Fight the Coronavirus Pandemic, *Scientific American.* https://www.scientificamerican.com/article/lessons-from-past-outbreaks-could-help-fight-the-coronavirus-pandemic1/

Gowland, R. (2006). The accidental risk assessment methodology for industries (ARAMIS)/ layer of protection analysis (LOPA) methodology : A step forward towards convergent practices in risk assessment ?, *130*, 307–310. https://doi.org/10.1016/j.jhazmat.2005.07.007

Hethcote, H. W. (1976). Qualitative analyses of communicable disease models. *Mathematical Biosciences*, *28*(3–4), 335–356. https://doi.org/10.1016/0025-5564(76)90132-2

Hethcote, H., Zhien, M., & Shengbing, L. (2002). Effects of quarantine in six endemic models for infectious diseases. *Mathematical Biosciences, 180(1–2), 141–160*. https://doi.org/10.1016/S0025-5564(02)00111-6

Hiorns, R. W., & MacDonald, N. (1982). Time Lags in Biological Models. *Journal of the Royal Statistical Society. Series A (General)*, *145*(1), 140. https://doi.org/10.2307/2981435

Hollingsworth, T. D. (2009). Controlling infectious disease outbreaks: Lessons from mathematical modelling. *Journal of Public Health Policy,* 30(3), 328 –341. https://doi.org/10.1057/jphp.2009.13

Iliuta, I., Larachi, F., Anfray, J., Dromard, N., & Schweich, D. (2007). Multicomponent multicompartment model for Fischer-Tropsch SCBR. *AIChE Journal*, *53*(8), 2062–2083. https://doi.org/10.1002/aic.11242

Institute of Medicine (2007). *Ethical and Legal Considerations in Mitigating Pandemic Disease*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/11917

Jamison, D. T., Gelband, H., Horton, S., Jha, P., Laxminarayan, R., Mock, C. N., & Nugent, R. (Eds.). (2017). *Disease Control Priorities, Third Edition (Volume 9): Improving Health and Reducing Poverty*. *Disease Control Priorities, Third Edition (Volume 9): Improving Health and Reducing Poverty*. The World Bank. https://doi.org/10.1596/978-1-4648-0527-1

Johns Hopkins University Center for Systems Science and Engineering, (2020). https://github.com/CSSEGISandData/COVID-19/blob/master/README.md

Jourdan, N., Neveux, T., Potier, O., Kanniche, M., Wicks, J., Nopens, Rehman, U., Moullec, Y. (2019). Compartmental Modelling in chemical engineering: A critical review. *Chemical Engineering Science*, *210*, 115196. https://doi.org/10.1016/j.ces.2019.115196

Kermack, W.O., McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *115*(772), 700–721. https://doi.org/10.1098/rspa.1927.0118

Khakzad, N., Khan, F., & Amyotte, P. (2012). Dynamic risk analysis using bow-tie approach. *Reliability Engineering and System Safety*, *104*, 36–44. https://doi.org/10.1016/j.ress.2012.04.003

Khakzad, N., Khan, F., & Amyotte, P. (2013). Safety Science Quantitative risk analysis of offshore drilling operations : A Bayesian approach. *Safety Science*, *57*, 108–117. https://doi.org/10.1016/j.ssci.2013.01.022

Khalid A, (2020, March 03). Coronavirus: 5 ways to put evidence into action during outbreaks like COVID-19, *The Conversation*. https://theconversation.com/coronavirus-5-ways-to-put-evidence-into-action-during-outbreaks-like-covid-19-131746

Khan, F. I., & Abbasi, S. A. (1998). Techniques and methodologies for risk analysis in chemical process industries, *11*, 261–277.

Khan, F. I., & Amyotte, P. R. (2005). I2SI : A comprehensive quantitative tool for inherent safety and cost evaluation, *18*, 310–326. https://doi.org/10.1016/j.jlp.2005.06.022

Khan, F., Rathnayaka, S., & Ahmed, S. (2015). Methods and models in process safety and risk management: Past, present and future. *Process Safety and Environmental Protection*, *98*, 116–147. https://doi.org/10.1016/j.psep.2015.07.005

Khoo, E. J., & Lantos, J. D. (2020). Lessons learned from the COVID-19 pandemic. *Acta Paediatrica*, *109*(7), 1323–1325. https://doi.org/10.1111/apa.15307

Koo, J. R., Cook, A. R., Park, M., Sun, Y., Sun, H., Lim, J. T., Tam, C., Dickens, B. L. (2020).

Interventions to mitigate early spread of SARS-CoV-2 in Singapore: a modelling study. *The Lancet Infectious Diseases, 20(6), P678–688*. https://doi.org/10.1016/S1473-3099(20)30162-6

Kopecki Lovelace J. B., Feuer W.,Dunn, N.,H., (2020, March 11). World Health Organization declares the coronavirus outbreak a global pandemic, CNBC, https://www.cnbc.com/2020/03/11/who-declares-the-coronavirus-outbreak-a-global-pandemic.html

Li, M. Y., Graef, J. R., Wang, L., & Karsai, J. (1999). Global dynamics of a SEIR model with varying total population size. *Mathematical Biosciences, 160(2), 191–213*. https://doi.org/10.1016/S0025-5564(99)00030-9

Lipsitch, M., Cohen, T., Cooper, B., Robins, J.M., Ma, S., James, L., Gopalakrishna, G., Chew, S.K., Tan, C.C., Samore, M.H., Fisman, D., Murray, M. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science*, *300*(5627), 1966–1970. https://doi.org/10.1126/science.1086616

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, *27*(2), 1–4. https://doi.org/10.1093/jtm/taaa021

Marhavilas, P. K., Koulouriotis, D., & Gemeni, V. (2011). Risk analysis and assessment methodologies in the work sites : On a review , classification and comparative study of the scientific literature of the period 2000 to 2009. *Journal of Loss Prevention in the Process Industries*, *24*(5), 477–523. https://doi.org/10.1016/j.jlp.2011.03.004

Markowski, A. S. (2007). exLOPA for explosion risks assessment, *142*(June 2006), 669–676. https://doi.org/10.1016/j.jhazmat.2006.06.070

Markowski, A. S., & Mannan, M. S. (2010). ExSys-LOPA for the chemical process industry. *Journal of Loss Prevention in the Process Industries*, *23*(6), 688–696. https://doi.org/10.1016/j.jlp.2010.05.011

Markowski, A. S., Mannan, M. S., & Bigoszewska, A. (2009). Fuzzy logic for process safety analysis. *Journal of Loss Prevention in the Process Industries*, *22*(6), 695–702. https://doi.org/10.1016/j.jlp.2008.11.011

Martcheva, M., & Castillo-Chavez, C. (2003). Diseases with chronic stage in a population with varying size. *Mathematical Biosciences, 182(1), 1-25*. https://doi.org/10.1016/S0025-

5564(02)00184-0

Word Health Organization, (2020). Coronavirus disease 2019 (COVID-19) Situation Report – 73. *WHO Bulletin*.

Safi, M. A., & Gumel, A. B. (2010). Global asymptotic dynamics of a model for quarantine and isolation. *Discrete and Continuous Dynamical Systems, 14(1), 209-231*. https://doi.org/10.3934/dcdsb.2010.14.209

Short, K. R., Kedzierska, K., & van de Sandt, C. E. (2018). Back to the Future: Lessons Learned From the 1918 Influenza Pandemic. *Frontiers in Cellular and Infection Microbiology*., 8, 343 https://doi.org/10.3389/fcimb.2018.00343

Srinivasan, R., & Natarajan, S. (2012). Developments in inherent safety : A review of the progress during 2001 – 2011 and opportunities ahead. *Process Safety and Environmental Protection*, *90*(5), 389–403. https://doi.org/10.1016/j.psep.2012.06.001

Sun, C., & Hsieh, Y. H. (2010). Global analysis of an SEIR model with varying population size and vaccination. *Applied Mathematical Modelling, 34(10), 2685-2697*. https://doi.org/10.1016/j.apm.2009.12.005

Tugnoli, A., Landucci, G., Salzano, E., & Cozzani, V. (2012). Journal of Loss Prevention in the Process Industries Supporting the selection of process and plant design options by Inherent Safety KPIs. *Journal of Loss Prevention in the Process Industries*, *25*(5), 830–842. https://doi.org/10.1016/j.jlp.2012.03.008

Vrábel, P., Van Der Lans, R. G. J. M., Cui, Y. Q., & Luyben, K. C. A. M. (1999). Compartment model approach: Mixing in large-scale aerated reactors with multiple impellers. *Chemical Engineering Research and Design*, *77*(4), 291–302. https://doi.org/10.1205/026387699526223

Wei, C., Rogers, W. J., & Mannan, M. S. (2008). Layer of protection analysis for reactive chemical risk assessment, *159*, 19–24. https://doi.org/10.1016/j.jhazmat.2008.06.105

Willey, R. J. (2014). Layer of protection analysis. In *Procedia Engineering*, 84, 12–22. https://doi.org/10.1016/j.proeng.2014.10.405

World Bank. 2020. Global Economic Prospects, June 2020. Washington, DC: World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/33748 License: CC BY 3.0 IGO.

World Health Organization (2020). Novel Coronavirus (2019-nCoV) Situation Report - 1. *WHO*

Bulletin

Worldometer (2020, June 21). https://www.worldometers.info/coronavirus/

Zhang, X. S., Pebody, R., Charlett, A., de Angelis, D., Birrell, P., Kang, H., Baguelin, M., Choi, Y. H. (2017). Estimating and modelling the transmissibility of Middle East Respiratory Syndrome Coronavirus during the 2015 outbreak in the Republic of Korea. *Influenza and Other Respiratory Viruses*, *11*(5), 434–444. https://doi.org/10.1111/irv.12467

Zhao, W., Buffo, A., Alopaeus, V., Han, B., & Louhi-Kultanen, M. (2017). Application of the compartmental model to the gas-liquid precipitation of CO_2 -$Ca(OH)_2$ aqueous system in a stirred tank. *AIChE Journal*, *63*(1), 378–386. https://doi.org/10.1002/aic.15567

Zhou, L., Miranda-Saksena, M., & Saksena, N. K. (2013). Viruses and neurodegeneration. *Virology Journal, 10,* 172. https://doi.org/10.1186/1743-422X-10-172

Chapter 7

Pandemic Risk Assessment and Management in a Bayesian Framework

Preface: This chapter extrapolates the goal of the preceding chapter of reliable assessment of the impact of a pandemic outbreak using advanced techniques. It presents a hierarchical Bayesian formulation for capturing variabilities in evolving conditions such as the advent of subsequent waves of a pandemic. The uncertainty in the reported pandemic data has been handled using the parameter sharing feature of the hierarchical Bayesian structure. It is also fitting our hypothesis of improving data-driven methods using science-based formulations of disease dynamics. This work can be mapped to the sub-objective of the thesis "devising semi-mechanistic models for assessing pandemic risk". The chapter has been submitted to the Risk Analysis journal.

Bibliographic citation and authorship statement: This chapter is a collaborative work among the Centre for Risk, Integrity and Safety Engineering (C-RISE), Faculty of Engineering and Applied Science, Memorial University of Newfoundland, Department of Process Engineering and Applied Science, Dalhousie University, and Department of Industrial Engineering, Dalhousie University. The first author (Md Alauddin) was the lead author and all the co-authors helped in developing the manuscript. The bibliographic citation and contributions of authors and co-authors are presented as follows.

Alauddin, M., Khan, F., Imtiaz, S., Ahmad, S., Amyotte, P., & Vanberkel, P. (2021). Pandemic Risk Assessment and Management in Bayesian Framework. *(submitted to Risk Analysis Journal).*

Md Alauddin: Formal Analysis, Methodology, Software; Investigation, Validation; Writing - Original Draft; Writing - Review & Editing

Faisal Khan: Conceptualization, Methodology, Writing - Review & Editing; Supervision; Project administration; Funding acquisition

Syed Imtiaz: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition

Salim Ahmed: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition

Paul Amyotte: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition

Peter Vanberkel: Methodology, Validation; Formal Analysis; Writing - Review & Editing; Supervision; Funding acquisition

Abstract: Credible assessment of a pandemic is critical in decision-making by healthcare providers, local and national government agencies, and international organizations. It facilitates a valuable lead time to policy-makers for efficient planning, resource allocation, and enforcing interventions. This work presents a Bayesian-based semi-mechanistic model for a short-term forecast of pandemic risk. We employed a hierarchical Bayesian network to learn the parameters of the susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) model to forecast the risk under evolving conditions due to imposed regulations, varied individual responses, and the advent of the multiple waves of a pandemic outbreak. The model has been validated for the risk forecast of the Coronavirus Disease of 2019 (COVID-19) using the benchmark models reported by the Center for Disease Prevention and Control and real-world data. We have also presented the impact analysis of COVID-19 pandemic risk using Bayesian inference. The proposed hierarchical Bayesian-based semi-mechanistic model (HBN-SEIQRD) resulted in accurate prediction with the lowest leave-one-out (LOO) cross-validation scores among the proposed Bayesian frameworks. The model is helpful in predicting the disease trajectory under evolving conditions due to government regulations, societal responses, and individual practices.

7.1 Introduction

Infectious diseases have been a persistent threat to human health for a long time. The Justinian Plague (541–542 AD), the Black Death (first outbreak in Europe in 1347), yellow fever in South America in the 16th century, the global influenza pandemic in 1918, severe acute respiratory syndrome (SARS), and HxNy influenza are some of the deadliest pandemics caused by zoonoses (Daszak, 2012). The case fatality rate (CFR) of infectious diseases has been significantly reduced due to recent developments in medical sciences; however, globalization, intense mobility, and complex networking, on the other hand, have increased the infectivity potential of diseases. Emergent diseases can turn into a pandemic impacting millions of people with loss of life, mental health, and severe economic impairment. The prevailing global Coronavirus disease (COVID-19) pandemic is an example of such zoonotic devastation.

Mathematical modeling helps understand disease dynamics and guides public health planning for effective control of an outbreak (Chowell, 2017). It helps in understanding the effectiveness of distinct administrative interventions such as lockdown, school and business closures, and a ban on social gatherings in reducing the spread of the disease. Modeling also provides scopes for the sensitivity analysis of the time frame of implementing and relaxing interventions in restraining the epidemic risk. The effects of early detection of infected cases, contact tracing, and quarantine of exposed cases have also been effectively captured in numerous models. Mathematical models have been recognized as valuable tools for clinical trials to examine the optimal cluster size and to validate of the trial's statistical analysis. Different types of mathematical methods and computational techniques such as differential equations, stochastic processes, statistical analysis, graph theory, artificial life, artificial society, computer simulation, and geographic information systems, have been developed to study epidemic diffusion in human societies. However, handling uncertainty is still a daunting task in pandemic risk management (Rozell, 2019).

Pandemic modeling comprises both aleatory uncertainty (caused by variability in population/data) and epistemic uncertainty (arising from a lack of knowledge of the phenomena). The distinct strains of the virus, modes of propagation (airborne or contact transmission), the existence or non-existence of asymptomatic spreading, and uncertainty in infectivity, rate of incubation, infection, and recovery all constitute uncertainty in epidemic modeling. Reported data of an outbreak are often erroneous due to various factors such as lack of systematic testing, the inherent delay between the date that an illness starts and the date the case is reported to public health authorities, and the accuracy of the test methods. Evolving conditions due to government regulations, societal responses, individual practices, and the outbreak of subsequent waves are other sources of uncertainty in predicting disease trajectory.

Cautionary/precautionary techniques help minimize fatality risk in a pandemic (Alauddin, Khan, Imtiaz, Ahmed, & Amyotte, 2021). For example, the lockdown of a geographical region until acquiring evidence of diminishing the disease's spread is a precautionary measure for reducing the fatality risk due to a pandemic outbreak. This, however, has severe socio-economic repercussions. Stringent lockdown and prolonged confinement can cause loss of employment, existential threats to organizations, and numerous mental health issues. Policy-makers should have a comprehensive picture of the scenarios realized due to distinct enforcement. This necessitates robust models for the credible forecast and assessment of inherent risk under varying conditions of a pandemic outbreak.

The Bayesian inference approach can help deal with uncertainty in pandemic risk management by incorporating conditional dependencies and updating prior beliefs with real-time information. We have proposed a hierarchical Bayesian model for parameter learning of the SEIQRD model to address temporal variability of the pandemic outbreak. The objective is to capture the variability of the parameters of each noticeable period while managing the uncertainty by enabling parameter sharing from respective periods of the outbreak. Our approach is based on tuning the model from local conditions and handling the temporal variability of the parameters caused by miscellaneous factors. We have also extended the concept of parameter sharing for estimating the risk due to multiple waves of a pandemic. The contributions of this paper include:

   i.   *Forecast of infected cases and mortality using a hierarchical Bayesian Network:* Pandemic risk forecasting is challenging due to evolving conditions caused by imposed regulations, varied individual responses, and the advent of multiple waves of an outbreak. To capture the temporal variability of the parameters of the SEIQRD mechanistic model, we have used Bayesian inference with a hierarchical structure. The variational Bayesian inference has been employed to generate a simplified proposal distribution of the complex

representation of the hierarchical model. Finally, the approximate posterior distributions of the parameters have been determined using the no U-turn sampler (NUTS) (Hoffman & Gelman, 2014). We have also examined other Bayesian network formulations such as pooling, and no pooling in handling temporal variabilities of a pandemic model.

ii. *Impact Analysis using Bayesian network:* The Bayesian inference network has been used to analyze the impact of distinct parameters in pandemic risk management. We have adopted relevance-based reasoning, which is based on the conditional probability distribution of the Bayesian network. We have studied the impact of the COVID-19 on peak hospitalization cases under three scenarios: no measures or few interventions, moderate interventions, and stricter interventions.

The remainder of the paper is organized as follows. Section 7.2 provides a brief survey of epidemiological models with a particular emphasis on compartmental models. It also presents a review of various approaches to parameter estimation of mechanistic models. Section 7.3 presents the mathematical models and the approaches to parameter learning in Bayesian frameworks. Risk assessment of COVID-19 for distinct scenarios is presented in Section 7.4 followed by the conclusion in Section 7.5.

7.2 Literature review

Infectious disease modeling can broadly be classified in two categories: statistical and mechanistic. Statistical approaches exploit data correlations to learn a functional dependence for predictions, while mechanistic models are based on physical laws such as population and/or transmission dynamics. In this section, we present an overview of mechanistic models for the dynamics and control of infectious diseases.

7.2.1 Mechanistic epidemiological models

Mechanistic models comprise compartmental, spatial, meta-population, network-based, and agent-based methods. The following sections present brief surveys of the studies on distinct mechanistic models.

7.2.1.1 *Compartmental epidemiological models*

Compartmental models are based on systems of ordinary differential equations that focus on the dynamic progression of a population through different epidemiological states (Chowell, 2017). The population is divided into several compartments, each having the same state of the epidemic. The SIR (susceptible, infected, recovered) and SEIR (susceptible, exposed, infected, recovered) are fundamental compartmental models. The SIR model presumes that the infected hosts become contagious immediately after exposure to an infected carrier, whereas the SEIR model considers the latency between exposures and infectious periods (Anderson & May, 1979; Hethcote, 1976; Hiorns & MacDonald, 1982; Kermack & McKendrick, 1927). Many extended compartmental models have been developed to take into account isolation, quarantine, and hospitalization (Alauddin et al., 2020; Arik et al., 2020; Giordano et al., 2020; Hu et al., 2020; Ivorra, Ferrández, Vela-Pérez, & Ramos, 2020; Legrand, Grais, Boelle, Valleron, & Flahault, 2007; Lin et al., 2020; Paiva, Afonso, de Oliveira, & Garcia, 2020; Subramanian, He, & Pascual, 2021).

The compartmental model is widely used for the timely assessment of epidemics. Many models have been developed to take into account the non-pharmaceutical interventions (NPIs) of isolation and quarantine of the exposed cases. Carvalho, da Silva, and Charret (2019) presented a comparative analysis of mechanical and chemical control methods in restraining a pandemic. The modeling of comprehensive interventions, treatment, and other control actions can be found in multiple studies (Alam, Kabir, & Tanimoto, 2020; Browne, Gulbudak, & Webb, 2015; Chowell & Kiskowski, 2016; Colizza, Barrat, Barthelemy, Valleron, & Vespignani, 2007; De Visscher, 2020; Fast, Mekaru, Brownstein, Postlethwaite,

& Markuzon, 2015; Jung, Lee, & Chowell, 2016; Lee, Lye, & Wilder-Smith, 2009; Rachah & Torres, 2016; Rizzo & Atti, 2008; Shen, Xiao, & Rong, 2015).

7.2.1.2 *Spatial epidemiological models*

Compartmental models presume well-mixed compartments that are hard to be observed in real life. These models do not consider social and spatial heterogeneity across geographic boundaries. Moreover, they cannot effectively harness the diverse data sources and surveillance aspects of an outbreak (Venkatramanan et al., 2018). According to Tobler's first law of geography, "everything is related to everything else, but near things are more related than distant things (Tobler, 1970)". Spatial structure and the position of the host population play a crucial role in the epidemic spread. Spatial models exploit the space property for random interactions in epidemic modeling. The most prominent advantage of spatial models is their visualizing power: spatial modeling maps are comprehensible tools for the general public. The heterogeneity among a larger population can be included in models using a population-based framework by dividing the large population into smaller groups of specific heterogeneity such as age, neighborhoods, and behavioral risk groups. Meta-population models assume uniform distribution of attributes across space within a subpopulation.

7.2.1.3 *Network-based epidemiological models*

Modeling of disease transmission in network-based models is achieved by describing more realistic contact patterns among individuals of the population (Lanzas & Chen, 2015). Network-based epidemiological models originated from mathematical graph theory, characterized by two components, namely the 'vertex' and the 'edge'. A vertex represents an individual, a group of people, a village, a city, or even an entire country whereas an edge represents the association between a pair of vertices. The interaction represented by such a graph could be directed or undirected (Martínez-López, Perez, & Sánchez-Vizcaíno, 2009). The determination of all possible interactions is cumbersome in a large

population. Nonetheless, mobile phones (Eagle, Pentland, & Lazer, 2008) and other global positioning system (GPS) data-loggers (Vazquez-Prokopec et al., 2013) have been crucial in establishing contact networks. The meta-population model is a widely adopted methodology to investigate the spatial-temporal transmission of infectious diseases. It presumes that the population is divided into diverse functional spatial subgroups called 'patches', such as cities, districts, villages, and schools. Differential equations are used within each subpopulation to describe disease dynamics. Meta-population models also incorporate spatial networks to characterize subpopulation movements and interactions (Apolloni, Poletto, Ramasco, Jensen, & Colizza, 2014; Colizza & Vespignani, 2008; Pell et al., 2016; Santermans et al., 2016).

7.2.1.4 Agent-based epidemiological models

The individual-based approach, also known as an agent-based model, mimics individual interactions. An individual is explicitly simulated with a set of characteristics, including spatial location, interaction preference, age, vaccination status, mobility, and stochastic behavior. Multi-agent simulations mimic the complexity of the system. A realistic synthetic population, social contact networks, and an effective disease model are the three basic components of agent-based epidemiological models (Eubank et al., 2004). Agent-based simulators include: MASON (Dunham, 2006), GeoGraph (Dibble & Feldmam, 2004), EpiModel (Jenness, Goodreau, & Morris, 2018), EpiSims (Mniszewski, Del Valle, Stroud, Riese, & Sydoriak, 2008), BioWar (Carley et al., 2006), and FluTE (Chao, Halloran, Obenchain, & Longini, 2010). Agent-based models require significant computational power and diverse data accessibility. They also yield sluggish responses in simulating diseases with low prevalence in a large population. Complex modeling frameworks such as meta-population, network-based, and individual-based models have been of increasing interest due to the availability of big data and increasingly sophisticated computational

techniques. However, the big data approaches often fail to reflect disease and social dynamics accurately. The performance of data-driven epidemiological models is further marred by poor data quality.

7.2.2 Parameter estimation of epidemiological models

Mechanistic model parameters can be estimated by least-squares fitting (Gan, Tan, Mo, Li, & Huang, 2020; Kim, Ko, Kim, & Jung, 2020; Romero-Severson, Ribeiro, & Castro, 2018), maximum likelihood estimation (Bretó, 2018; Chowell, Nishiura, & Bettencourt, 2007; Kao & Eisenberg, 2018; White & Pagano, 2008; Wu & Riley, 2016), and approximate Bayesian computation (Almutiry & Deardon, 2020; Brown, Porter, Oleson, & Hinman, 2018; Chandra, 2020; Kypraios, Neal, & Prangle, 2017; McKinley et al., 2018; Minter & Retkute, 2019; Neal, 2019; Neal & Huang, 2015; Price, Bean, Ross, & Tuke, 2016). Epidemiological modeling has been extensively studied using deterministic approaches (Carcione et al., 2020; Clancy & Piunovskiy, 2005; House & Keeling, 2008; Kumar et al., 2020; Lange, 2016; Lipsitch et al., 2003; Okyere et al., 2020; Otoo et al., 2020; Pollicott et al., 2012; Rivers et al., 2014; Sharkey, 2011; Zhou et al., 2014) and stochastic approaches (Alharthi, Kypraios, & O'Neill, 2019; Allen, 2017; Birrell, de Angelis, & Presanis, 2018; Bjørnstad, Finkenstädt, & Grenfell, 2002; Chao et al., 2010; Engbert, Rabe, Kliegl, & Reich, 2021; Fintzi, Cui, Wakefield, & Minin, 2017; Khan, Hussain, Zahri, Zaman, & Humphries, 2020; Kypraios et al., 2017; Lekone & Finkenstädt, 2006; O'Neill, 2002; Ponciano & Capistrán, 2011; Shangguan, Liu, Wang, & Tan, 2021; Taylor, Dushoff, & Weitz, 2016; Wang, Ji, Bi, & Liu, 2020). Pollicott et al. (2012) reconstructed a time-dependent transmission rate for the SIR model. Camacho et al. (2015) modeled a time-varying transmission parameter of SEIR by the Wiener process (also known as standard Brownian motion). Cauchemez et al. (2008) employed a stochastic framework and Markov chain Monte Carlo (MCMC) to recover time-dependent transmission rate and other model parameters. Alauddin et al. (2020) proposed a semi-mechanistic artificial neural network-based SEIQRD model for capturing time varying parameters. Google Cloud researchers integrated the time series machine

learning approach with a compartmental model to develop Google Cloud AI (Arik et al., 2020). Population genetic inference using coalescent models has been exploited in the parameter estimation of compartmental models (Dearlove & Wilson, 2013; Popinga, Vaughan, Stadler, & Drummond, 2014). Many studies have investigated branching processes to estimate the basic reproduction number (R0) and the final size of the epidemic (Allen, 2015; Ball & Donnelly, 1993; Tritch & Allen, 2018).

Evolutionary computing approaches have also been explored for estimating parameters of deterministic and stochastic epidemiological models. For example, a genetic algorithm was employed for the parameter estimation of several epidemic diseases such as Cholera (Akman & Schaefer, 2015), Malaria (Davis et al., 2019), SARS (Isa Irawan & Amiroch, 2015; Yan & Zou, 2008), HIV-AIDS (Pedroso-Rodriguez, Marrero, & De Arazoza, 2003), and COVID-19 (Anđelić, Šegota, Lorencin, Mrzljak, & Car, 2021; Ding et al., 2021; Hosseini, Ghafoor, Sadiq, Guizani, & Emrouznejad, 2020; Monteiro, Gandini, & Schimit, 2020). Eastman et al. (2021) used differential evolution to fit their model to the epidemiological data of COVID-19. Similarly, swarm intelligence has been used for fitting epidemic models (Akman, Akman, & Schaefer, 2018; He, Peng, & Sun, 2020; Hosseini et al., 2020).

Uncertainty is a key aspect of pandemic modeling. Ambiguity in the mechanism of disease transmission, randomness in the rate of incubation, infection, and recovery periods, and biological transmutations of the disease-causing elements, constitute uncertainty in epidemic modeling. This is further complicated by distinct mitigative measures and public responses while the outbreak unfolds (Dehning et al., 2020). Uncertainty is a major challenge for policymakers, as it impacts reliability of model outcomes. Sensitivity analysis helps to identify key factors for the spread of disease by quantifying the effects of uncertainties; Monte Carlo methods, Fourier amplitude sensitivity tests, Latin hypercube, Bayesian estimation (Arriola & Hyman, 2009) are common tools for sensitivity analysis.

The predictive ability of a model can be improved by reducing the variability in the forecasts. Xu et al. (2016) proposed a Bayesian non-parametric method for stochastic epidemic models using a Gaussian process. Oliveira et al. (2020) employed a Bayesian approach to the SIR model with correction for under-reporting in the analysis of COVID-19 cases (Manevski, Ružić Gorenjec, Kejžar, & Blagus, 2020). Polo et al. (2020) used a Bayesian model to study the spatio-temporal effect on health and fatality caused by the COVID-19. Bayesian inference was used for studying the time dependence of effective growth rate of new infections (Das & Tiwari, 2021; Dehning et al., 2020). Farah et al. (2014) developed a Bayesian framework for parameter estimation of a computationally expensive dynamic epidemic model using time series epidemic data. Dehning et al. (2020) applied Bayesian inference based on Monte Carlo sampling to characterize the change points realized due to distinct mitigative measures. Smirnova et al.(2019) introduced the reconstruction of nonparametric time-dependent transmission rates by projecting onto a finite subspace spanned by Legendre polynomials to obtain an accurate approximation of model parameters.

Numerous methods in the Bayesian framework have been devised to treat data inconsistency and under-reported cases (Deardon et al., 2010; Gibson, Reich, & Sheldon, 2020; Russell et al., 2020; Sharmin, Glass, Viennet, & Harley, 2018; Taghizadeh, Karimi, & Heitzinger, 2020). Li et al. (2018) compared distinct models of varying complexity on three MCMC platforms: JAGS (Plummer, 2003), NIMBLE (Lawson, 2020), and Stan (Carpenter et al., 2017). Lytras et al.(2019) introduced FluHMM based on a hidden Markov model fitted in a Bayesian framework. Azmon et al. (2014) investigated the impact of delays and under-reporting by using a Bayesian semiparametric approach with penalized splines. Markov switching model has been employed to determine the epidemic and non-epidemic periods from surveillance data (Amorós, Conesa, López-Quílez, & Martinez-Beneito, 2020; Martínez-Beneito, Conesa, López-Quílez, & López-Maside, 2008). Dureau et al.(2013) employed an adjusted adaptive particle

Markov chain Monte Carlo algorithm for handling parameter uncertainty. Bayesian inference using a Markov chain Monte Carlo (MCMC) approach for parameter estimation in epidemic models has been reported in several inquests (Almutiry & Deardon, 2020; Brown et al., 2018; Cotta, Naveira-cotta, & Magal, 2020; Deardon et al., 2010; M. Li et al., 2018; Marwa, Mwalili, & Mbalawata, 2019; Osei, Duker, & Stein, 2012; Taghizadeh et al., 2020; Touloupou, Finkenstädt, & Spencer, 2020).

Lee and Mallick (2020) employed a hierarchical Bayesian approach for estimating COVID-19 spread by borrowing information from global data. However, the epidemic disease trajectory depends on demographics, economic status, degree of compliance of the population and societal impact, comorbidities, overall risk environment, and country vulnerability to biological threats. We propose a hierarchical Bayesian mechanistic model based on the SEIQRD model for pandemic risk assessment. Our approach is based on tuning the model from local conditions to capture variabilities due to evolving conditions caused by imposed regulations, varied individual responses, and the advent of multiple waves of an outbreak. The outbreak trajectory of a geographical area is divided in several periods, each having distinct parameters. Despite the parameters being different for each period, they share similarities due to the hierarchical representation of the entire period. The next section outlines the development of the proposed model.

7.3   The model development

This section presents a hierarchical Bayesian model of disease transmission following an outbreak. Section 7.3.2 outlines the overall procedure for the pandemic risk assessment and management using the proposed Bayesian framework. The basic understanding of distinct modules has been presented in the preliminary section.

7.3.1 Preliminaries

In this section, we will briefly describe the SEIQRD epidemiological model followed by the parameter estimation under varying conditions using a hierarchical Bayesian network. This section also presents variational inference and the MCMC sampling which have been used for the parameter estimation of the proposed model.

7.3.1.1 The SEIQRD model

The susceptible, exposed, infected, quarantined, recovered, deceased model captures the effects of asymptomatic transmission, hospitalization, and quarantine on disease spread (Fig. 7.1). The mathematical formulations of the SEIQRD model are presented in Eqs.7.1-7.7, where 'a', 'b', 'c', and 'e' respectively denote the rates of contagion, incubation, infection, and recovery. 'N' represents the population of the geographical area, 'd' the rate of hospitalization/quarantine after being symptomatic, 'φ_1' the fraction of symptomatic infections, and 'φ_2' the fraction of quarantined/hospitalized population resulting in mortality. The details of the model can be found in Alauddin et al. (2020).
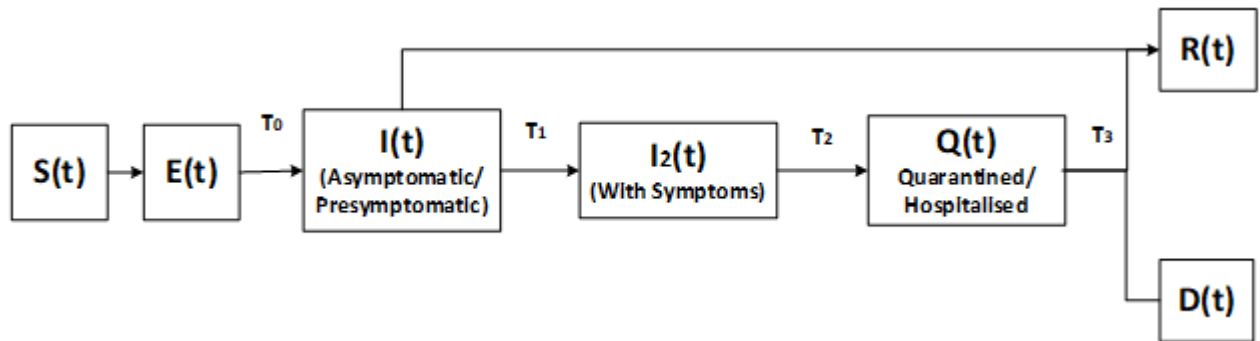
Fig. 7.1: Schematic representation of the SEIQRD model for infectious disease transmission (T_0: incubation period, T_1: infection period, T_2: duration between case detection and quarantine/hospitalization, T_3: recovery period)

$$\frac{dS(t)}{dt} = -\frac{aS(t)I(t)}{N} \tag{7.1}$$

$$\frac{dE(t)}{dt} = \frac{aS(t)I(t)}{N} - bE(t) \tag{7.2}$$

$$\frac{dI(t)}{dt} = bE(t) - cI(t) \tag{7.3}$$

$$\frac{dI_2(t)}{dt} = \varphi_1 cI(t) - dI_2(t) \qquad (7.4)$$

$$\frac{dQ(t)}{dt} = dI_2(t) - eQ(t) \qquad (7.5)$$

$$\frac{dR(t)}{dt} = (1 - \varphi_1)cI(t) + (1 - \varphi_2)eQ(t) \qquad (7.6)$$

$$\frac{dD(t)}{dt} = \varphi_2 eQ(t) \qquad (7.7)$$

The model can be written in the next-generation form (Yang, 2014) as $\frac{dF}{dx} = F(X) - V(X)$, where

$F(X)$ denotes the rate of appearance of new infections in a compartment, and $V(X)$ gives the transfer of

individuals between distinct compartments. The basic reproduction number R0 is given by the spectral

radius of the matrix FV^{-1} (Appendix B). From the next-generation method, if the disease-free reproduction

number Rdf < 1, then the disease-free equilibrium point is locally asymptotically stable, and if Rdf > 1,

then it is unstable. The model is based on the following assumptions:

(i) Natural births, as well as deaths due to other reasons during the study period, are not considered in counting the total population.

(ii) Recovered people are immune to further viral attacks during the short span of analysis. This assumption is clinically proven by many studies for various contagious viral attacks (Zhou et al., 2013; Short, Kedzierska & Van de Sandt, 2018).

(iii) There is no infectivity (transmission) in hospitalized or self-isolated compartments. This is possible if the advisory is strictly followed during treatment and isolation.

7.3.1.2 Hierarchical Bayesian network

Bayesian network (BN) is an effective tool for safety and risk analysis. It can incorporate multi-state

variables, conditional dependencies, and real-time information to update prior beliefs. A Bayesian

network is a directed acyclic graph (DAG) comprising nodes and arcs. The node represents the probability

distribution of a random variable, while an arc determines the probabilistic relationship between two

connected variables. BN employs the Bayes theorem to generate posteriors based on updating the prior occurrence probabilities of events using extant information, called evidence. A BN works by propagating belief in the entire network and is termed as a Bayesian Belief Network (BBN) (Mallick & Imtiaz, 2013).

Bayesian inference allows for more readily dealing with and interpreting uncertainty and easier incorporation of prior beliefs. There are three approaches to modeling in a Bayesian framework: (i) pooling of measurements, (ii) unpooled measurements, and (iii) partial pooling or hierarchical models. In pooling, the parameters are estimated for the entire period using one comprehensive model, $y = f(\theta)$. This will not effectively capture variations in the number of cases due to distinct regulations and public response over different phases of the outbreak. Moreover, it is not capable of modeling the multiple peaks of a pandemic. Separate models for each period $y_i = f(\theta_i)$ can be employed in the unpooled measurements to account for this. However, these predictions suffer from considerable uncertainty due to the lack of large datasets for each phase. Both of these limitations are addressed by hierarchical models where the parameters of the individual periods θ_i are different for each period as in the unpooled case, but they share coefficients with the common prior distributions θ as illustrated by Fig. 7.2. This can also facilitate estimating the controlling parameters in the subsequent waves of a pandemic outbreak. Limited and sparse data can be efficiently used through the hierarchical structure of distinct levels (Lim et al., 2018).

Eq. 7.8 provides the posteriors of the parameters under given evidence using the Bayes rule.

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y,\theta)d\theta} \qquad (7.8)$$

The expression $\int p(y,\theta)d\theta$ has no closed-form solution over a high-dimensional space. This integral in Bayes rule is approximated by Markov chain Monte Carlo (MCMC) sampling. MCMC sampling is slow

and flounder when the model complexity increases. Thus, we have employed variational inference based on approximate distribution instead of finding the real posterior distribution.
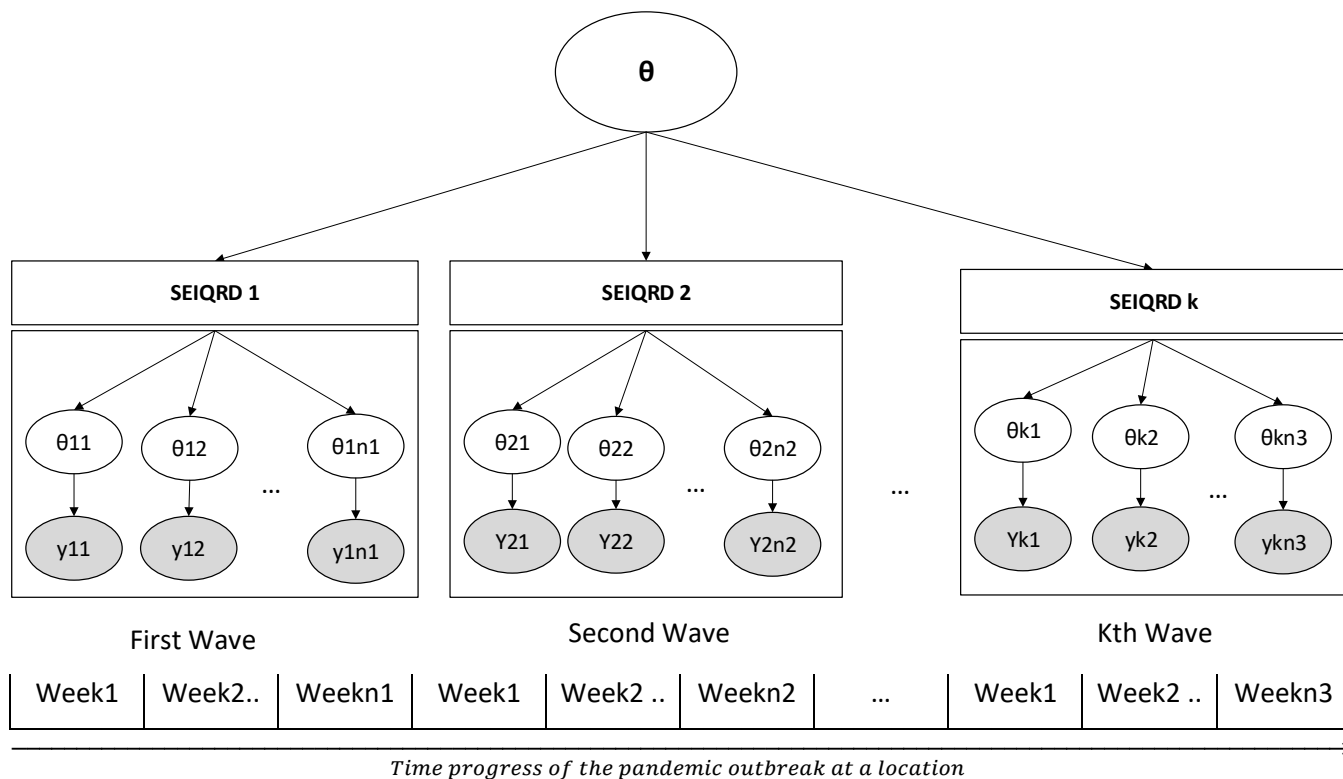
Fig. 7.2: Modeling of multiple waves of a pandemic using a hierarchical Bayesian network

7.3.1.3 Variational Inference

The main idea of variational inference is to approximate the real posteriors with a tractable family of distributions. In this space, we search for an optimal distribution $q*$ that minimizes a certain measure of dissimilarity (Eq. 7.9).

$$q^* = argmin_{q \epsilon Q} \, f\big(q(.), p(.|y)\big) \qquad (7.9)$$

Where $f(\,)$ is a measure of dissimilarity. We have employed the KL-divergence for the dissimilarity. The KL-divergence can be represented by Eqs 7.10-17.11.

$$KL(p\|q) = \int p(x) ln \left(\frac{p(x)}{q(x)}\right) dx \qquad (7.10)$$

$$KL(q\|(p(\theta|y))) = \int q(\theta) ln\left(\frac{q(\theta)}{(p(\theta|y))}\right) d\theta \qquad (7.11a)$$

$$= \int q(\theta) ln\left(\frac{q(\theta)p(y)}{p(\theta,y)}\right) d\theta \qquad (7.11b)$$

$$= \log p(y) - \int q(\theta) \log \frac{p(y|\theta)p(\theta)}{q(\theta)} d\theta \qquad (7.11c)$$

Since $p(y)$ is fixed, we only need to consider the second term, which is also known as the *evidence lower bound (ELBO)*. Thus, to minimize KL-divergence, we need to maximize the ELBO.

Variational inference turns approximate posterior inference into a computationally efficient optimization problem. Though it is a promising method, developing a variational inference algorithm still requires tedious model-specific derivations and implementation. We have employed automatic differentiation variational inference (ADVI) that automates the process of deriving scalable variational inference algorithms as illustrated in Fig. 7.3 (Kucukelbir, Blei, Gelman, Ranganath, & Tran, 2017).

Transform the latent variables of the model into a common space

Choose a variational approximation in the common space

Stochastic approximation

Posterior distribution
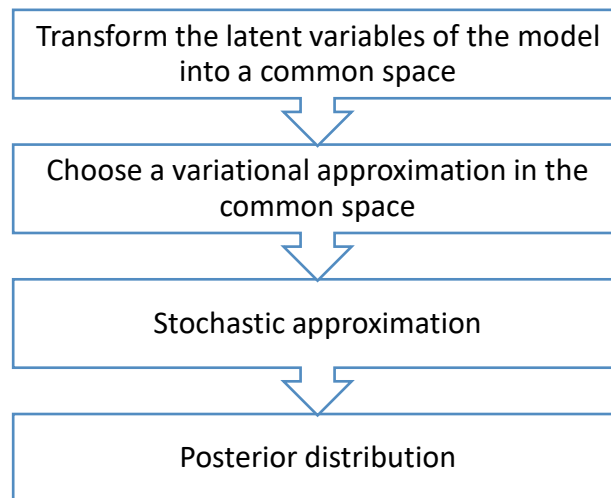
Fig. 7.3: Steps for automatic differentiation variational inference (ADVI)

7.3.1.4 *Markov chain Monte Carlo sampling*

Markov chain Monte Carlo sampling (MCMC) is a method of sampling a high-dimensional probability distribution. It is based on the Markov chain mechanism that exhibits a memory-lessness property where

the next sample is dependent only on the current state (Eq. 7.12). Eq. 7.13 presents a detailed balance equation which is a sufficient, but not necessary, condition to ensure that a particular p(x) is the desired invariant distribution (Andrieu, De Freitas, Doucet, & Jordan, 2003). P(x) and T(x) denote the distribution and transition matrix. The samplers are designed for effective convergence while maintaining the balance condition.

$$P\left(x^{(i)}\middle|x^{(i-1)}, x^{(i-2)}, \dots\dots, x^{(1)}\right) = T\left(x^{(i)}\middle|x^{(i-1)}\right) \tag{7.12}$$

$$P\left(x^{(i)}\right)T\left(x^{(i)}\middle|x^{(i-1)}\right) = P\left(x^{(i-1)}\right)T\left(x^{(i-1)}\middle|x^{(i)}\right) \tag{7.13}$$

Metropolis-Hastings and Gibbs Sampling are two well-known MCMC sampling techniques. The fundamental problem with Metropolis-Hastings and Gibbs sampling is that they are too random and often tend to get trapped in narrow regions of moderately complex targets. The Hamiltonian Monte Carlo (HMC) method can handle random walk behavior and sensitivity to correlated parameters with a fewer samples to obtain an efficient target distribution. However, HMC requires specifications of steps for its simulated paths. The step number determines how long the path continues before a new flick is made in a new random direction. This results in similar samples when converged in too few steps, and highly dissimilar samples with many-step convergence.

No U-Turn Sampler (NUTS) is an approach for adaptively finding the optimal number of steps. NUTS is based on a recursive algorithm to build a set of likely candidate points that spans a large proportion of the target distribution. The NUTS algorithm tries to determine when the path starts to turn around and results in more effective samples per iterations than Gibbs or Metropolis-Hasting for complex posteriors (Hoffman & Gelman, 2014; Nishio & Arakawa, 2019). HMC and NUTS take advantage of gradient information from the likelihood to achieve much faster convergence than traditional sampling methods, especially for larger models. NUTS also has several self-tuning strategies for adaptively setting the tunable parameters of Hamiltonian Monte Carlo. We have employed NUTS for sampling in our model to estimate

the parameters of the distinct SEIQRD models of hierarchical Bayesian frameworks presented in Fig. 7.4 and Fig. 7.5.

7.3.2 The pandemic risk management using hierarchical Bayesian network

Fig. 4 presents the flowchart for pandemic risk assessment and impact analysis in a Bayesian framework. We have employed a hierarchical structural based Bayesian framework to estimate the epidemiological parameters of the SEIQRD model using MCMC sampling. The uncertainty in the reported data has been handled using the parameter sharing feature of the hierarchical Bayesian structure. We used variational inference to approximate the real posterior distribution of parameters such as the rate of incubation, infection, and recovery, and other parameters of the SEIQRD model (Fig. 4B). These approximations serve as initializing point for the MCMC sampling. Finally, No U-Turn Sampler (NUTS) has been used for Markov Chain Monte Carlo sampling from the approximated probability distributions. The posterior parameters of the SEIQRD mechanistic models help in estimating pandemic risk in terms of infections and fatalities. This also helps in determining the number of hospitalization cases, which are critical in pandemic risk management.

Data generated from the SEIQRD model has been used to perform impact analysis of the model parameters. This has been carried out using relevance-based reasoning based on the conditional probability distribution of the Bayesian network (Arora et al., 2019). We have discretized the infection risk into three categories: low, medium, and high depending upon the respective peak hospitalization cases. The parameters under study are also discretized in a distinct range. For instance, the impact of distinct interventions could be studied by dividing it into three categories: little or no intervention, moderate interventions, and stricter interventions such as curfew. The incubation, infection, and recovery periods have also been analyzed for three levels. The presence and absence of the virus (SARS-CoV-2)

has also been modeled as a Boolean function in determining the impact on pandemic risk. Eq 7.14 and Eq. 7.15 present the joint probability and conditional probability distributions of the system.

$$P(I, T0, T1, T3, R0, V) = P(I|T0, T1, T3, V)P(T0|T1, T3, V)P(T1|T3, V)P(T3|V)P(V) \qquad (7.14)$$

$$P(I|T0, T1, T3, V) = \frac{P(T0, T1, T3, V|I)}{P(T0, T1, T3, V)}P(I) \qquad (7.15)$$

Where, $I, T0, T1, T3, R0, and\ V$ respectively denote the risk of infection, incubation period, infection period, recovery period, reproduction number, and the presence of the Coronavirus.

We present a hierarchical Bayesian-based approach to the susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) model of infectious disease spread. Various models in the Bayesian framework based on pooling and no pooling of data have also been examined for risk forecasting under temporal variations due to imposed regulations, varied individual responses, and the advent of multiple waves of a pandemic outbreak. The following section presents the procedure of the study and the model outcomes on risk forecasts of COVID-19. The model has been validated using benchmark models and on real-world data.
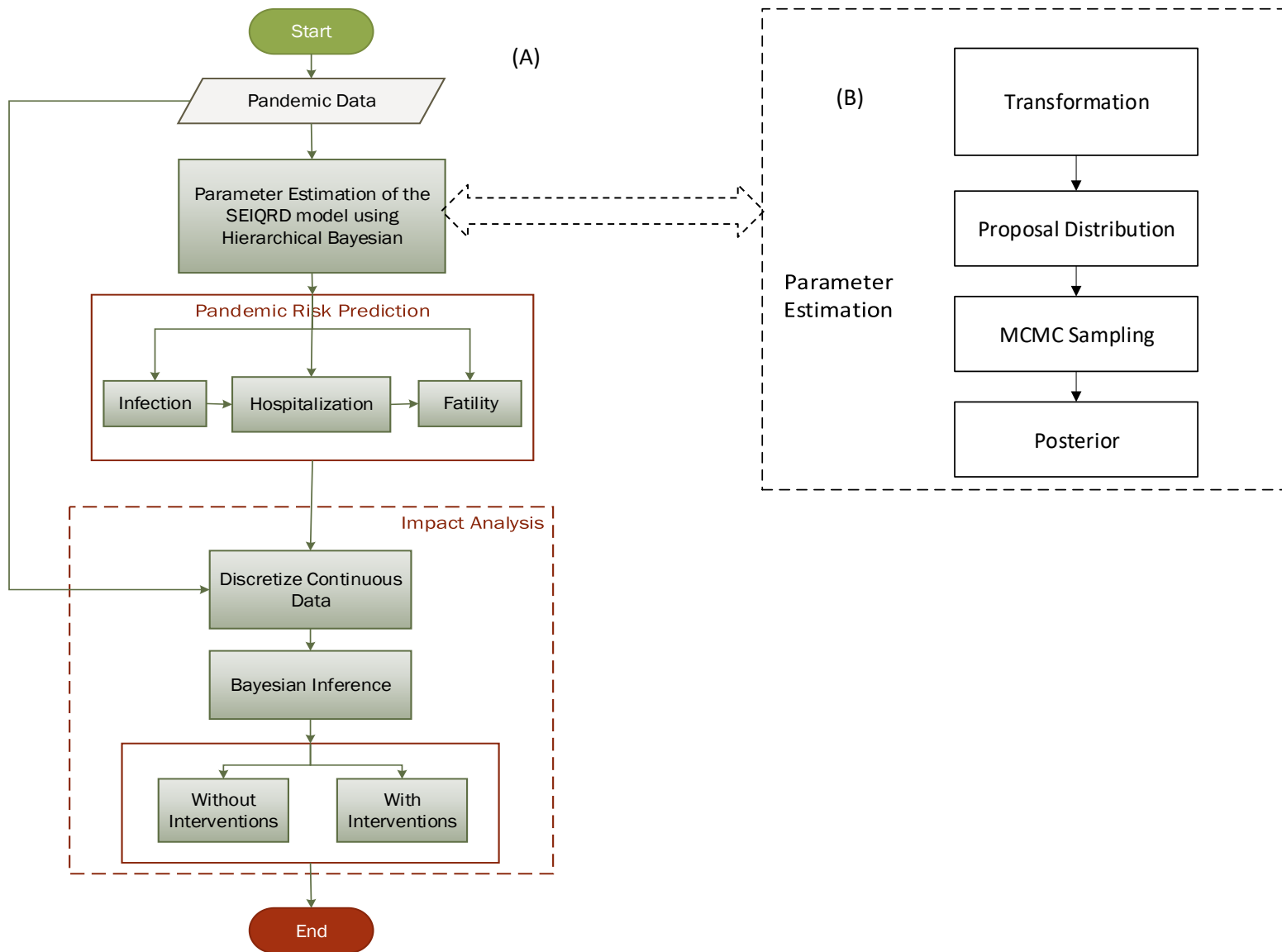
Fig. 7.4: Pandemic risk assessment and impact analysis in a hierarchical Bayesian framework (A) Outline of the process, (B) Parameter estimation using variational Bayesian inference

7.4   Risk forecast of COVID-19

We have examined the proposed models for forecasting pandemic impacts due to the first and second waves of COVID-19 in Ontario, Canada and Arizona, USA. The procedure could be summarized in three groups: estimating model parameters, forecasting short-term scenarios, and analyzing impacts. The task has been carried out as follows-

Step 1: Accessing and preprocessing of pandemic data

We received data from March 17, 2020 to Jan 2021 for Ontario and Arizona. The data have been accessed using Johns Hopkins University Center for Systems Science and Engineering (JHU CSSE) and CDC COVID-19 Data Tracker (https://covid.cdc.gov/covid-data-tracker/#datatracker). The pandemic data were processed for smoothening and removing spikes and noise. We divided the data in two groups corresponding to wave 1 and wave 2. Each group was further divided into subgroups consisting of 15 days periods. For example the processes data for Ontario $D^{(i,j)}$ is presented in Eq. 7.16, where the first index 'i' represents the wave, and the second index 'j' represents the intended period in the ith wave.

$$D^{(i,j)} = \begin{cases} D^{(1,1)} \ from \ Mar \ 17 \ to \ Mar \ 31, 2020 \\ D^{(1,2)} \ from \ Apr \ 01 \ to \ Apr \ 15, 2020 \\ D^{(1,3)} \ from \ Apr \ 16 \ to \ Apr \ 30, 2020 \\ \qquad\qquad \ldots \\ \qquad\qquad \ldots \\ D^{(1,10)} \ from \ Jul \ 30 \ to \ Aug \ 13, 2020 \\ D^{(2,1)} \ from \ Aug \ 14 \ to \ Aug \ 28, 2020 \\ \qquad\qquad \ldots \\ \qquad\qquad \ldots \\ D^{(2,9)} \ from \ Dec \ 12 \ to \ Dec \ 26, 2020 \\ D^{(2,10)} \ from \ Dec \ 27 \ to \ Jan \ 10, 2021 \end{cases} \qquad (7.16)$$

Step 2: Initialization of priors of the model parameters

The informative prior make Bayesian analysis effective in dealing with uncertainty and inconsistency of the pandemic data. We selected informative priors on initial model parameters based on the values commonly reported in the literature. However, we also provided a sufficient exploration range by providing adequate spread parameters (Table 7.1). We have used a bounded half-normal prior for the

basic reproduction number with a lower bound of 1.0 and an average value of 2.9. We used a high value of spread parameters to accommodate distinct values reported in various studies. The basic reproduction number 2.0–2.5 has been initially reported by the WHO-China Joint Mission on Coronavirus Disease 2019. The Australian health authorities reported R_0 in the range 2.6–2.7, which was subsequently revised to 2.5–3.5 (Chang et al., 2020). For example, a median $R_0 = 3.4$ (CI [2.4, 4.7]) was used in a model of the COVID-19 spread in Germany (Dehning et al., 2020). Flax et al. (2020) estimated the initial R to be 3.8 (2.4–5.6) from a study on 11 European countries. A review by Liu et al, (2020) presented the R-value ranging from 1.4 to 6.49, with a mean of 3.28 and a median of 2.79. The higher exploration range can also aid the variability caused due to mutations, non-pharmaceutical interventions and public responses.

The average incubation period, infection period, and recovery period were assigned lognormal priors with values of 5.5, 5.1, and 11.5 days, respectively (Liu et al., 2020; Bi et al, 2020). Asymptomatic and fatality fractions were assigned as bounded normal priors with respective average values of 0.15 and 0.01. Several studies adopted the log-normal distribution to capture the tail in the parameters of pandemic models (Ejima et al, 2020). We selected log-normal distributions of incubation infection, and recovery periods. Thus, the log-normal priors for the incubation period T0~LogNormal[log(5.5),1], infectious period T1~LogNormal[log(5.1),1], and recovery periods T3~LogNormal[log(11.5),1] have been employed in this study.

Step 3: Initialization of the Markov chains through variational inference

We have employed variational inference for generating effective starting points for MCMC sampling. This can lead to accelerated convergence of the MCMC. We have used PYMC3 for generating the inference, the source code has been provided in Appendix D.

Table 7.1: Priors of the hierarchical Bayesian-based SEIQRD model (HBN-SEIQRD)

PARAMETERS	PARAMETERS NAME	VARIABLES	PRIOR DISTRIBUTIONS
	Reproduction Number	R_0	BoundedNormal(lower=1)[2.9, 2]
Hyper Parameters	Incubation Period	T_0	LogNormal[5.5, 1]
	Infection Period	T_1	LogNormal[5.1, 1]
	Recovery Period	T_3	LogNormal[11.5, 1]
	Symptomatic fraction	φ_1	Bounded Normal(lower=0.1, upper=1)[0.8, 0.2]
	Mortality fraction	φ_2	Bounded Normal(lower=0, upper=0.1)[0.8, 0.2]
Parameters for i^{th} wave	Reproduction Number	$R_0[i, j]$	BoundedNormal(lower=1)[mu=R_0, shape=number of 15 days periods in i^{th} wave]
	Incubation Period	$T_0[i, j]$	LogNormal[mu=T_0, shape=number of 15 days periods in i^{th} wave]
	Infection Period	$T_1[i, j]$	LogNormal[mu=T_1, shape=number of 15 days periods in i^{th} wave]
	Recovery Period	$T_3[i, j]$	LogNormal[mu=T_3, shape=number of 15 days periods in i^{th} wave]
	Symptomatic fraction	$\varphi_1[i, j]$	Bounded Normal(lower=0.1, upper=1)[mu= φ_1, shape=number of 15 days periods in i^{th} wave]
	Mortality fraction	$\varphi_2[i, j]$	Bounded Normal(lower=0, upper=0.1)[mu= φ_1, shape=number of 15 days periods in i^{th} wave]

Step 4: The MCMC sampling

The MCMC sampling has been used to generate the posterior distributions of the model parameters. It can be performed using different samplers based on the nature of the model. We have employed the No_U turn sampler. After sampling a fraction of the initial samples also known as the Burn-in phase are discarded to retain effective samples for posterior analysis.

Step 5: Forecast using MCMC samples

The posterior parameters can be used for forecasting in the intended time period. The Bayesian-based forecast provides a credible interval based on the distribution of posterior parameters.

Step 6: Impact analysis:

The Hospitalization data corresponding to the SEIQRD model was generated using the hyper parameters. We developed peak hospitalization based on distinct sampled parameters. The parameters and the hospitalization cases were discretized to find the studied range. For example, we discretized the reproduction number in three categories, greater than_equal to 2 (R_2_up), between 1 and 2 (R_1_2) and_less_than unity (R_bellow_1). The pandemic risk can be categorized as low risk (if hospitalization peaks< 10000), moderate (10000< if hospitalization peaks < 50000), and high risk (if hospitalization peaks> 50000) based on the availability of the healthcare facility (Table 7.2). Finally, the impact analysis is carried out performed using the conditional probability distributions of the system (Eq. 7.14-7.15).

Table 7.2: Discretization of dependent variable and parameters

Reproduction number	Discretized Reproduction number	Risk	Discretized Risk
5.1	R_2_up	1000	Low_Risk
1.5	R_1_2	100000	High_Risk
0.8	R_bellow_1	80000	High_Risk
1.2	R_1_2	30000	Moderate_Risk
1.6	R_1_2	500	Low_Risk
..	..	..	..
2.5	R_2_up	20000	Moderate_Risk
2.1	R_2_up	150000	High_Risk

Fig. 7.5 presents the digraph of the hierarchical Bayesian network for risk assessment of the COVID-19 pandemic. Here, "cases" represent the newly infected cases and cumulative fatality due to the COVID-19 pandemic. We have modeled two waves of the outbreak, each having different parameters of the SEIQRD model. Waves 1 and 2 are divided into 10 and 8 biweekly periods, respectively. Each period has different parameters governed by priors at the top layer that enable sharing of information between the two waves and among distinct periods of a particular wave. PyMC3 (https://docs.pymc.io/) has been used for variational inference, MCMC sampling, and forecasting of scenarios while GeNIe 2.2 (https://www.bayesfusion.com/) has been employed for the inference studies in the impact analysis (Section 4.3). The next section presents the parameter estimation of the proposed HBN-SEIQRD model.

First wave   Second wave
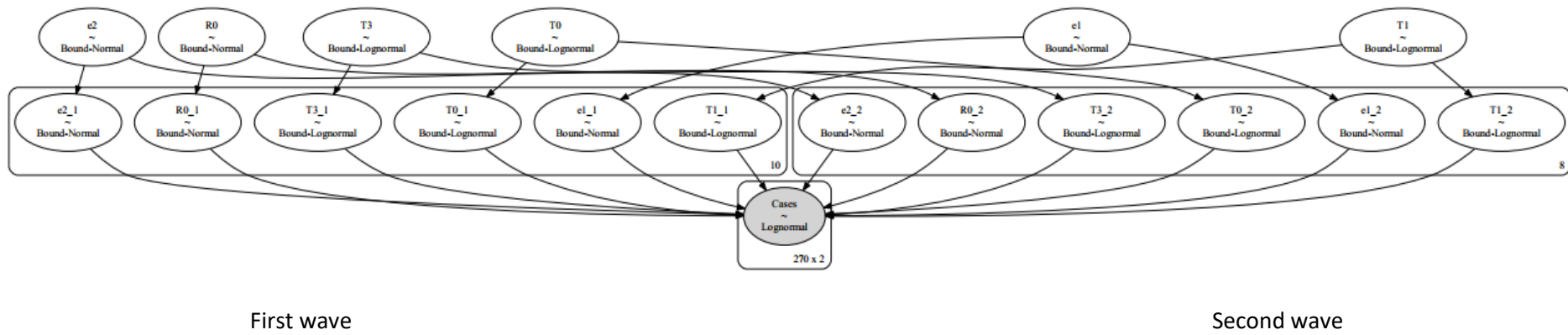
Fig. 7.5: Digraph of the hierarchical Bayesian network for risk assessment of COVID-19 pandemic.

7.4.1 Parameter estimation of the hierarchical Bayesian-based SEIQRD model

The parameters of the hierarchical Bayesian-based SEIQRD model have been estimated using Markov chain Monte Carlo sampling. We have employed automatic differentiation variational inference (ADVI) to generate effective starting points for the accelerated convergence of MCMC sampling. Variational inference is most suited for analyzing big data, however, it can also be used for initializing points for the MCMC sampling. The *adaptive gradient optimization* method has been used to minimize the negative *evidence lower bound (NELBO)* that minimizes the KL-divergence (Fig. 7.6). The NELBO converged to 4260.5 with 500 iterations. We adopted the NUTS sampling strategy to extract higher effective samples per iterations for generating the posterior distributions of the model parameters and hyper-parameters. We have withdrawn 10,000 samples from the posterior distribution for the traces (Appendix C). The stability of the algorithm is ensured by burning the initial samples.
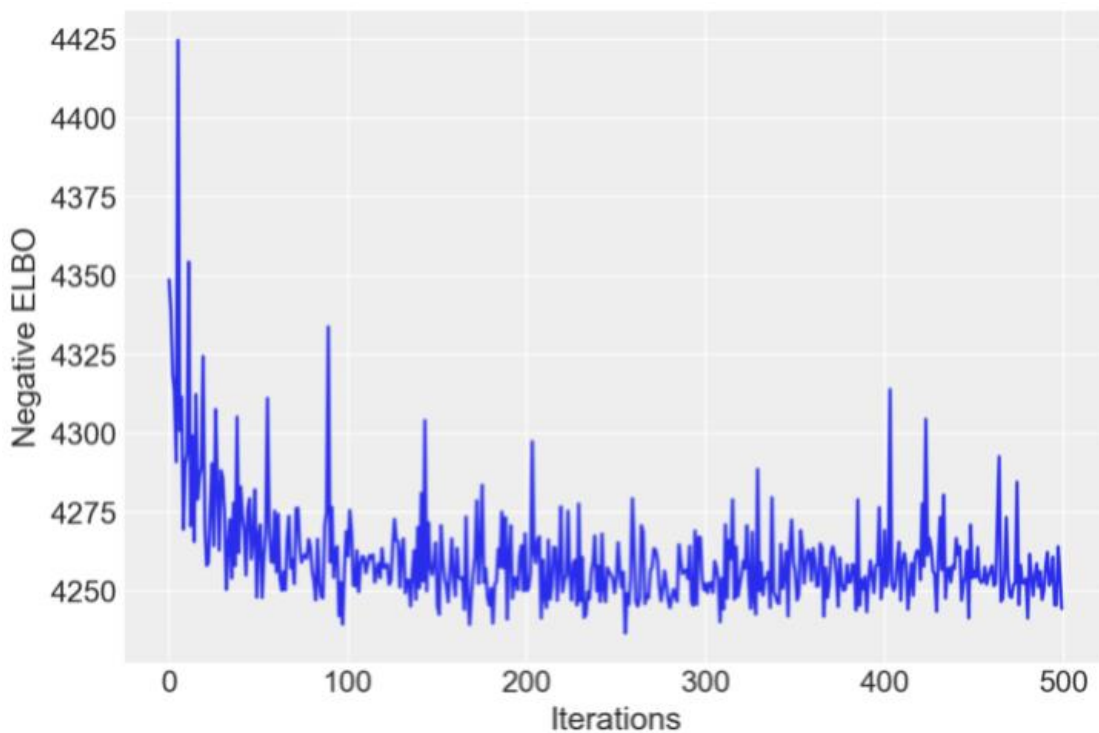
Fig. 7.6: The ELBO profile with iterations (*Average loss = 4260.5)*

Bayesian analysis provides posterior distribution that best represents parameters according to the data and the model. This distribution is a balance of the prior and the likelihood. Fig 7.7 presents the posterior of the hyper parameters of the proposed HBN-SEIQRD model. Table 7.3 presents the effective reproduction number of COVID-19 in selected studied periods during the first wave of the COVID-19 outbreak at Ontario in terms of the mean and the highest probability density (HPD). The HPD summarizes marginal posterior distributions by tabulating 100(1 – significance level) % posterior credible intervals for the parameters of interest (Turkkan & Pham-Gia, 1997).
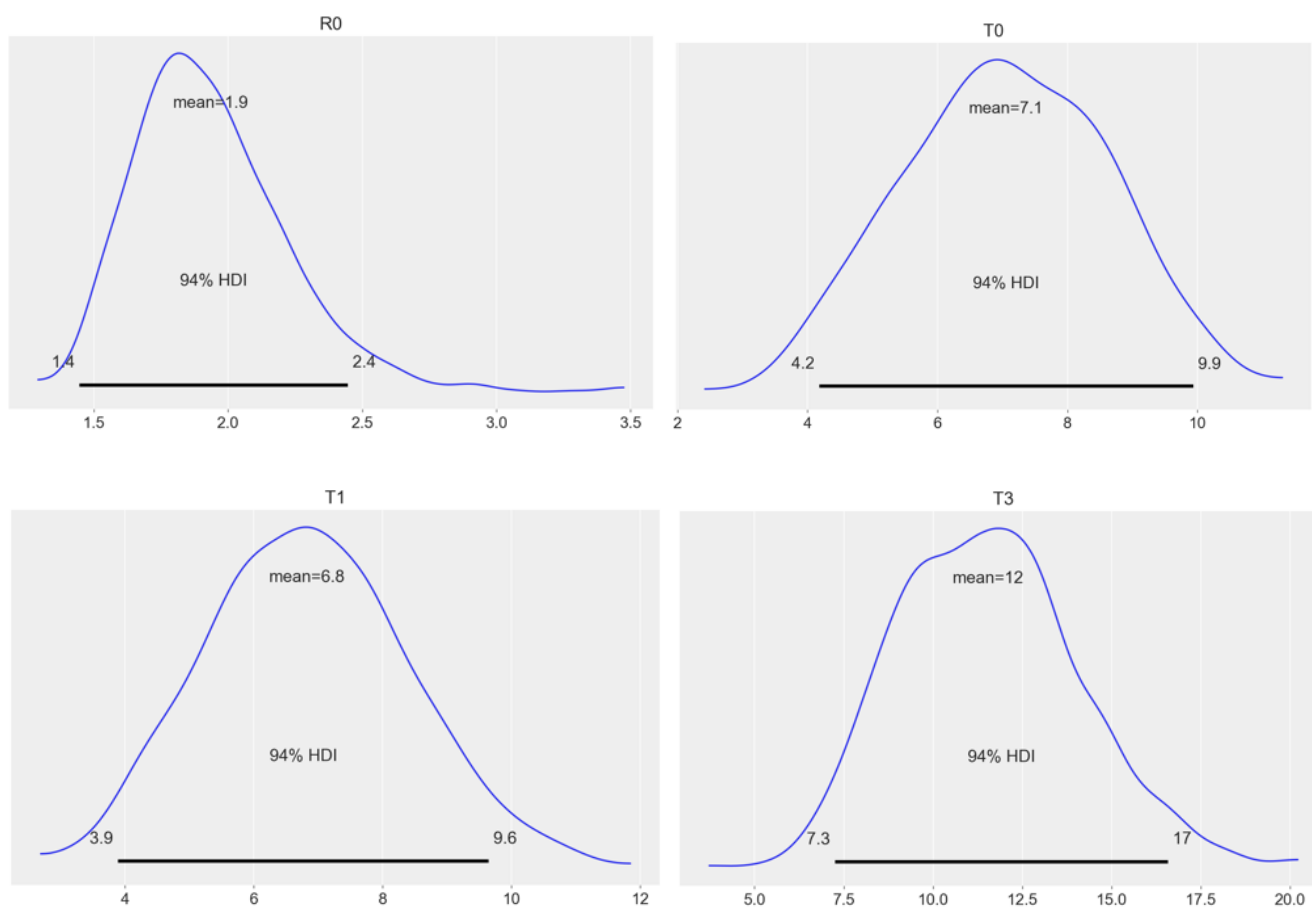
Fig. 7.7: The posterior distribution of the hyper parameters of the hierarchical Bayesian-based SEIQRD model for the first wave of the COVID-19 outbreak at Ontario.

Table 7.3: The mean value and the credible interval of the effective reproduction number of COVID-19 in selected studied periods during the first wave of the COVID-19 outbreak at Ontario.

Reproduction number for distinct study periods during the first wave	Mean Value	94% HDI
R-value for the Ist period-R1_1	2.5	1.3 to 3.9
R-value for the 2nd period-R1_2	1.9	1.2 to 2.7
R-value for the 3rd period-R1_3	1.8	1.2 to 2.8
R-value for the 4th period-R1_4	1.9	1.2 to 2.8
R-value for the 5th period-R1_5	2.1	1.3 to 3.2
R-value for the 6th period-R1_6	1.8	1.1 to 2.6
R-value for the 7th period-R1_7	1.8	1.1 to 2.8
R-value for the 8th period-R1_8	1.8	1.2 to 2.6
R-value for the 9th period-R1_9	1.7	1.1 to 2.6
R-value for the 10th period-R1_10	1.8	1.2 to 2.7

An important step when using MCMC for Bayesian inference is to perform diagnostic checks to reassure the user that the parameter space has been well-explored by the Markov chain and that the chain has converged to the equilibrium distribution. However, a converged model is not guaranteed to be a good model. Traceplots represent the value of the chain for each step in the sampling process. Ideally, all chains should trace the same distribution represented by the overlap in the plot. There should not be any detectable pattern, and none of the traces should flatten out. From the traceplots (Appendix C), we can observe that the sampling process converged indicating the actual parameter distributions of the model. The second component of model checking, the goodness of fit, is used to check the internal validity of the model by comparing predictions from the model to the data used to fit the model. The next section presents the forecast by the proposed hierarchical Bayesian-based SEIQRD model. We validated our findings for the pandemic risk forecast for the second wave with actual reported cases and numerous models presented by the Centre for Disease Prevention (https://covid.cdc.gov/covid-data-tracker/#forecasting_weeklydeaths) as presented in Appendix E.

7.4.2 Forecast of COVID-19 pandemic impacts the hierarchical Bayesian-based SEIQRD model

Fig. 7.8 presents a one-month ahead forecast of cumulative death due to COVID-19 in Ontario using the

hierarchical Bayesian-based SEIQRD model. Fig. 7.9 presents a one-month-ahead forecast of cumulative

death due to COVID-19's second wave in Arizona. We can see that the proposed hierarchical Bayesian

model (HBN-SEIQRD) forecast agrees with those models listed by the Centre for Disease Prevention

(Appendix E).
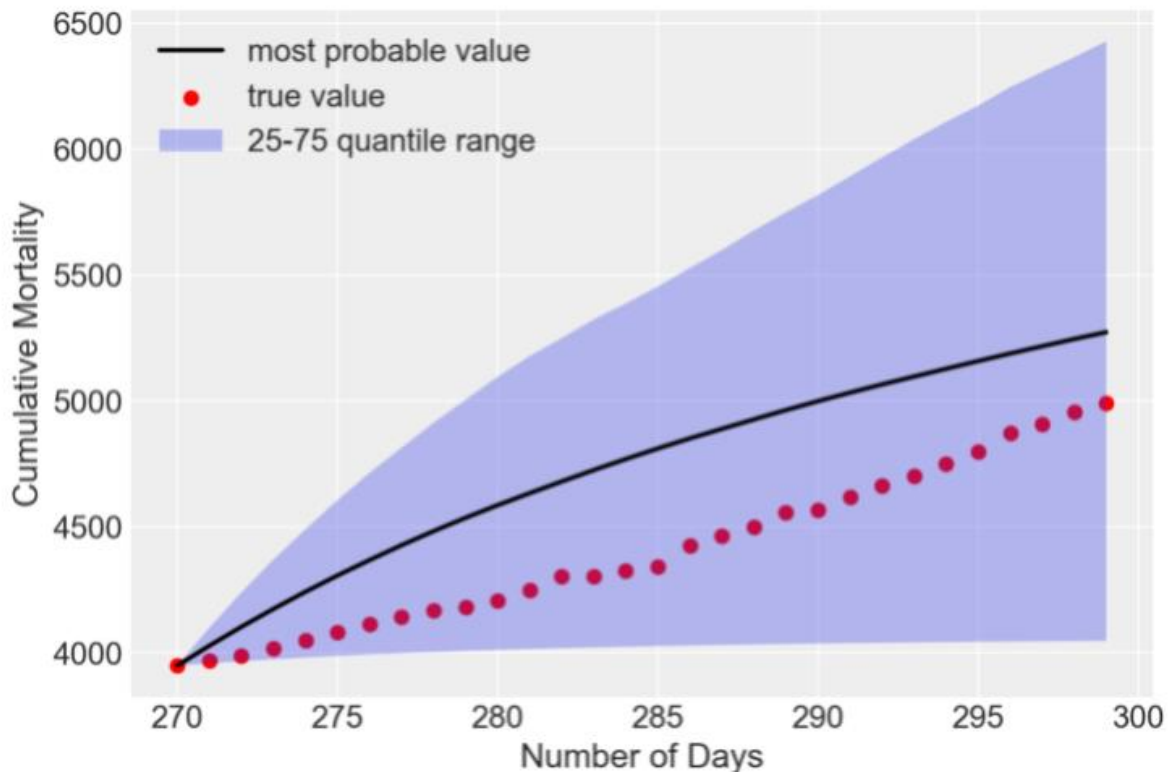
Fig. 7.8: Forecast of cumulative death due to COVID-19 in Ontario using hierarchical Bayesian network

We have also explored other possible Bayesian network formulations for risk forecasting using pooling

(Fig. 7.10a), and no pooling (Fig. 7.10b) of the data from both waves of the outbreak. With pooling (Fig.

7.10a), we have assumed that there is one comprehensive model for the entire duration of each of the

waves; i.e., pooling is with respect to the waves of the outbreak. Fig. 7.10b presents the model where data

sharing among different biweekly periods is not carried out from the same or different waves. Each

biweekly time horizon independently estimates the parameters of the model based on the corresponding data. We have examined 10 and 8 biweekly periods for the first and the second waves, as shown in Fig 7.10 a-b.

Table 7.4 presents the comparative performance of the models. The rank presents the relative ranking of the models starting from 0 (best model) to the number of models. The leave-one-out cross-validation (LOO) is an estimate of the out-of-sample predictive fit. It estimates the accuracy of the predictive distribution $p(\tilde{y}_i|y)$ by splitting samples in training and validation sets (Vehtari, Gelman, & Gabry, 2017). The LOO is asymptotically equivalent to the Akaike Information Criterion (AIC) of the frequentist's approach (Watanabe, 2010). However, the LOO effectively handles the uncertainty in the parameters which are ignored by the multivariate normal distributional assumption of the AIC. Lower LOO scores represent a better match between model and data. The term d_loo is the relative difference between the values of LOO of the top-ranked model with other models. Among these Bayesian formalism, the hierarchical Bayesian model (HBN-SEIQRD) scores well with reference to the leave-one-out cross-validation. Its performance relative to the pooled model could be improved by taking more samples in each individual period.
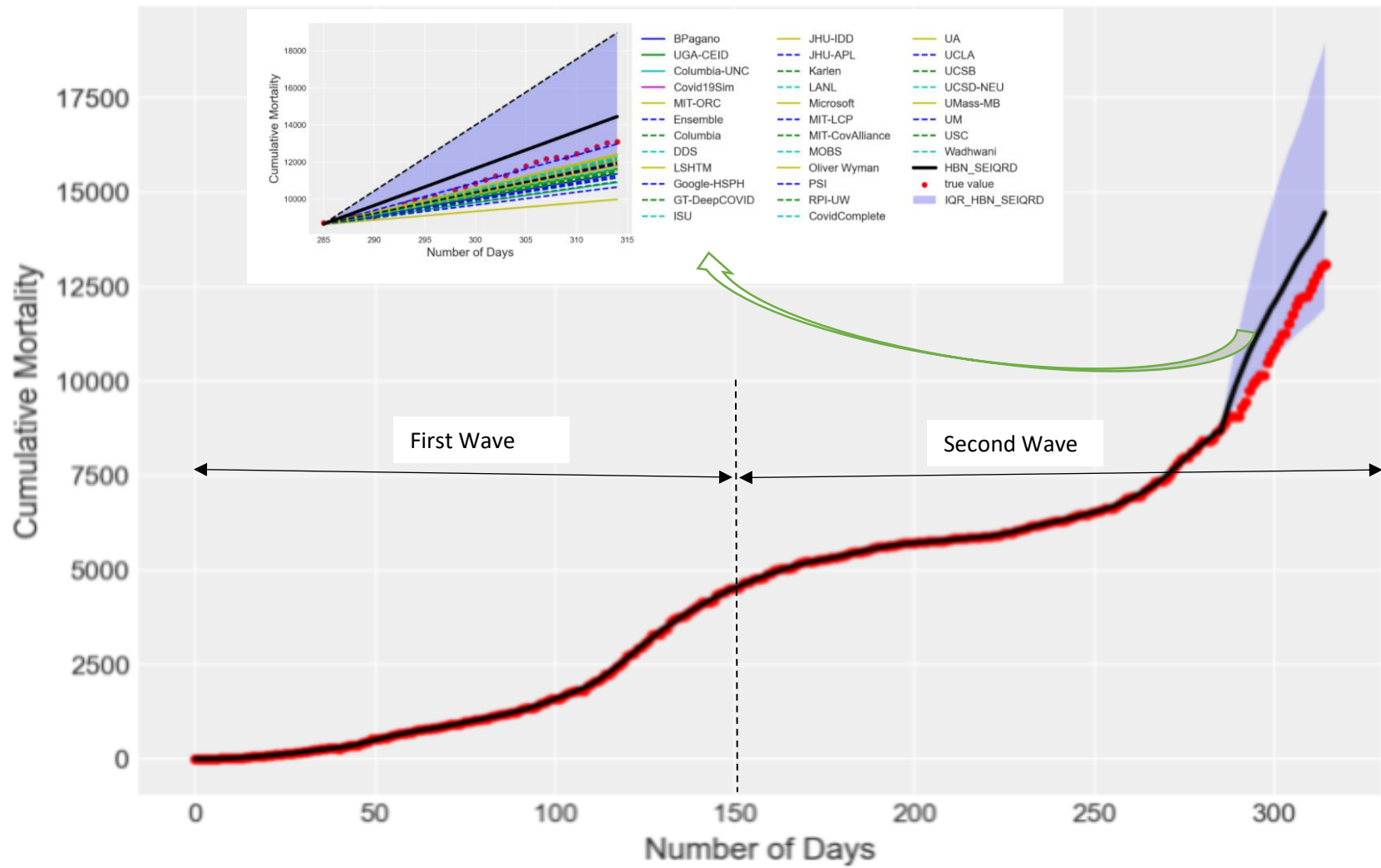
Fig. 7.9: A one-month ahead forecast of cumulative death due to COVID-19 in Arizona
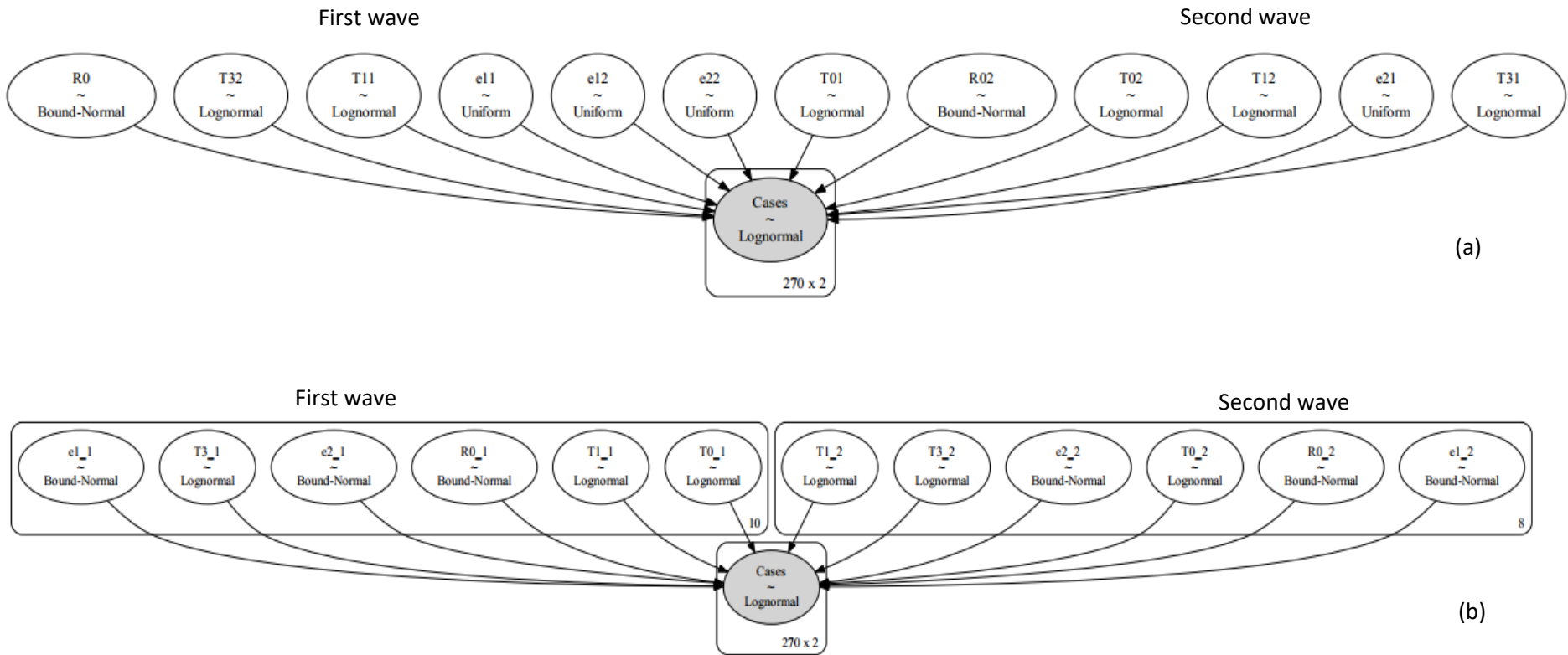
Fig. 7.10: Bayesian network formulations for the COVID-19 risk forecast using the SEIQRD models: (a) Pooling, (b) No pooling

Table 7.4: Comparative sampler statistics of test models of Bayesian frameworks for the posterior parameters of the SEIQRD model for the COVID-19 outbreak in Ontario

Models	Rank	Relative difference between the values of LOO (d_loo)
HBN-SEIQRD	0	0.00
Pooled-SEIQRD	1	12.8
Unpooled-SEIQRD	2	37.32

The proposed hierarchical model would perform well if there are no abrupt changes. For instance, super spreading events in a specific period could distort the model outcomes. Similarly, in studying the distinct waves using the hierarchical representation, the model would perform well if the nature (i.e., peak, growth rate) of all the subsequent waves of a location are alike. However, we noticed that subsequent waves, especially second waves, were distinctly different at many places. Nonetheless, that could also be captured by incorporating the correlation of subsequent waves from other places. The scenario is also different if vaccination is introduced in a specific studied period or wave. The next section presents the impact analysis of the model parameters using the relevance-based reasoning of Bayesian inference.

4.3 Impact Analysis using Bayesian network

Fig. 7.11 presents the overall inference network for COVID-19 pandemic risk with and without interventions. The main idea behind interventions is to reduce the exposure factor and slow down disease transmission. During a pandemic outbreak, a locality can be protected from the disease-causing element (e.g., SARS_CoV_2 in COVID-19) by border closure and/or effective screening at entry points. Many studies reported that border closures could have helped limit viral transmission in the pandemic's early days; their impact diminishes with community transmission of the disease. The benefits of border closures were also momentary when they were not accompanied by other measures such as testing, contact tracing,

and quarantining to prevent local transmission (Chinazzi et al., 2020; Hossain et al., 2020; Mallapaty, 2021).

The immunity gained through natural (herd) immunity or vaccination is the ultimate protection from pandemic risk. However, this takes several months following an outbreak. Many studies proposed temporal and spatial segregation measures to restrain a pandemic. Government interventions such as lockdown, school and business closures, and restricting large gatherings help in slowing down the spread of a pandemic by reducing exposure. Corporations and employers can assist in controlling the risk by transforming operational formats, such as enabling home delivery services, working from home, and switching to a virtual mode for meetings (Alauddin et al., 2021). They can also help in reducing exposure by installing shields at cash and other counters, separate entrances and exits, and signage for safe distancing. However, these interventions cannot be imposed for a longer duration due to high incurred costs and other associated risks. Stringent lockdown and prolonged confinement can cause neuropsychiatric problems, psychological disorders, and weakened immune systems. Interventions should be time-limited, reconsidered, and revisited regularly for effective handling of health and socio-economic impacts.

The effective infectious period could be reduced significantly by adopting contact tracing, increasing testing capacity, and devising methods for minimizing testing time. Contact tracing is characterized by identifying and monitoring each person who has been in contact with an infected person. The main advantages are that it can identify potentially infected individuals before severe symptoms emerge, and if conducted sufficiently quickly, can prevent onward transmission from the secondary cases. It is particularly effective when the latent period is long. Contact tracing is simplest and most effective when cases are symptomatic. However, evidence of other modes of transmission was also reported in COVID-19 transmission, including symptomatic (Li et al., 2020), pre-symptomatic (Nissen et al., 2021; Tong et

al., 2020; Wei et al., 2020), asymptomatic (Bai et al., 2020), and environmental (Ferretti et al., 2020). Thus other measures need to be enforced. Individual practices and societal responses are central to the effectiveness of these measures. Self-imposed measures such as wearing a mask in public places, voluntary social distancing, and handwashing are vital in preventing subsequent waves of an outbreak. Providing sophisticated treatment to infected people is vital for their safe recovery. Existing healthcare facilities might need to be extended to meet the demands of treating a large number of infected cases. A comprehensive review of interventions has been presented in several references (Alauddin et. al., 2021; Davies et. al., 2020; Ferguson et. al., 2005; Giordano et. al., 2020; Li et al., 2020; Rayner Brown et al., 2021).
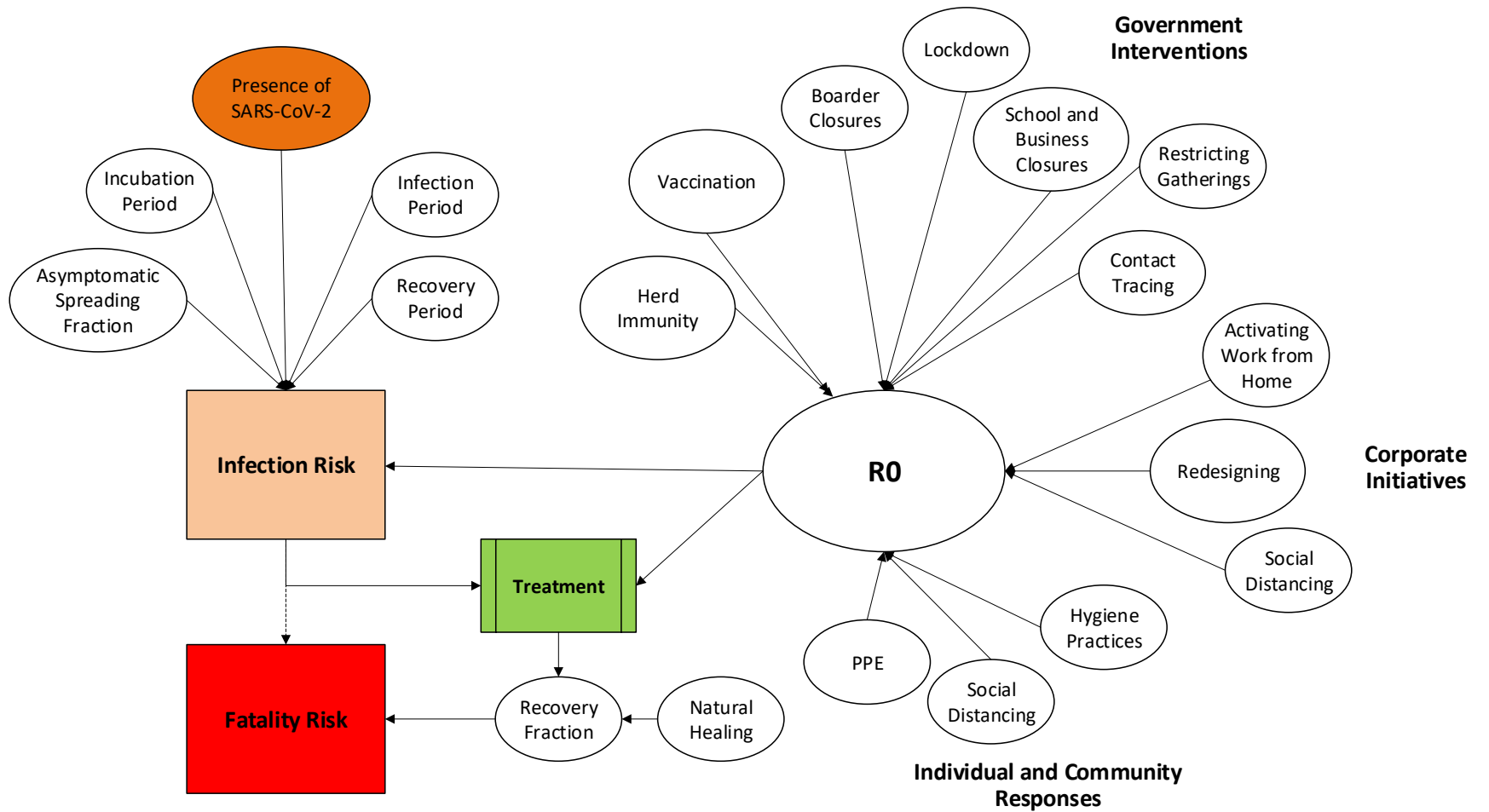
Fig. 7.11: Inference network for COVID-19 pandemic risk with interventions

Fig. 7.12 presents a simplified version of the inference network for the COVID-19 pandemic for studying the number of hospitalization cases. We have discretized the infection risk into three categories, low, medium, and high depending upon the respective peak hospitalization cases of < 10,000, 10,000-50,000, and > 50,000. We have assumed that Ontario's initial health care system has 10000 acute care beds as reported in (Barrett et al., 2020). We categorize the peak hospitalization cases of 10,000-50,000 a day as moderate risk assuming that a large fraction of the mild cases would be treated at home. We have studied the impact of distinct interventions by dividing the effective reproduction number into three categories: R> 2 (little or no intervention), 1 <R< 2 (moderate interventions), and R <1 (stricter interventions such as curfew). The incubation, infection, and recovery periods have also been analyzed for three levels.

After imparting evidence of the presence of the Coronavirus, average incubation period of 3-6 days, average infection period of 3-6 days, and average recovery period of 10-14 days, we can observe that the risk is very high if no action is enforced (Fig. 7.12A). We want to have a low risk even though the virus is present in the locality. This could be possible if there are moderate and stricter interventions (Fig. 7.12B-C).
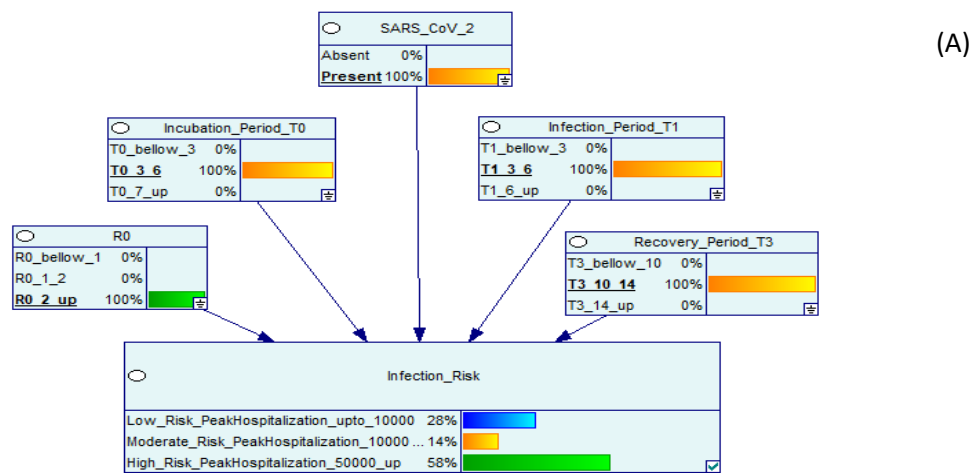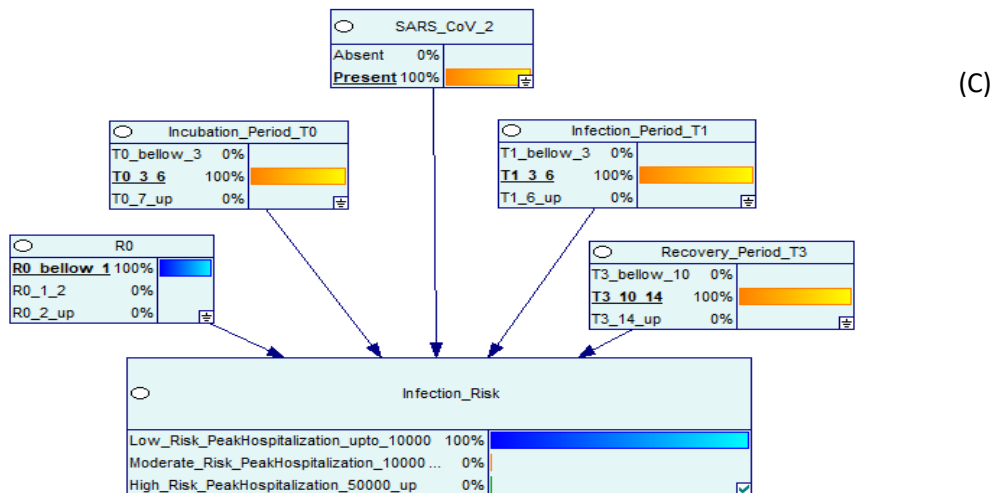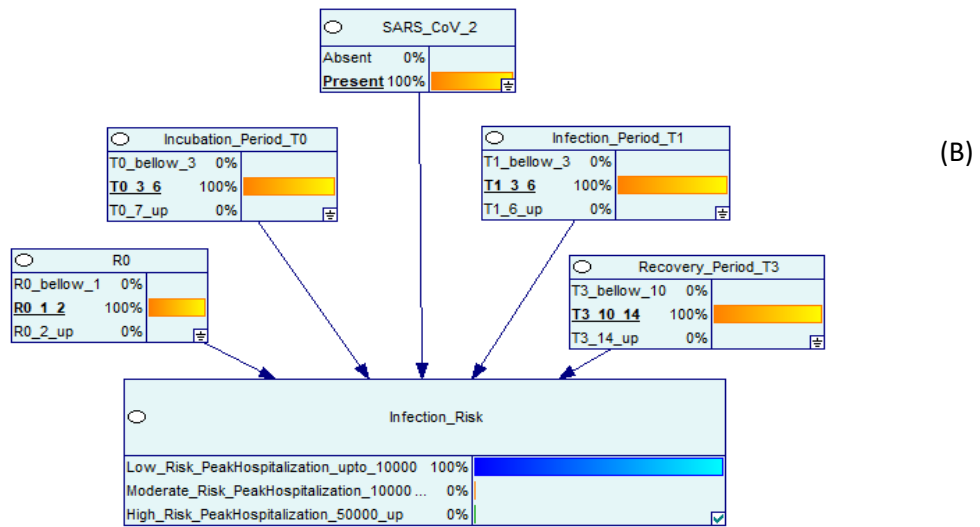
Fig. 7.12: Simplified Inference network for COVID-19 pandemic risk: (A) No measures or little intervention (B) Moderate interventions, and (c) Stricter inventions for bringing the effective reproduction number < 1.

5. Conclusions

We have proposed a semi-mechanistic hierarchical Bayesian-based SEIQRD model (HBN-SEIQRD model) for pandemic risk management. The resulting model predictions account for uncertainty in the mechanism of disease transmission, uncertainty in the severity of the disease, and randomness in incubation, infection, and recovery periods. The model can account for changing scenarios such as government regulations, societal responses, individual practices, and the advent of multiple waves of a pandemic. We obtained the posterior distributions for the model parameters instead of confidence intervals, which are commonly reported with frequentist methods. The Bayesian approach facilitated using the preliminary knowledge of the parameters as priors in estimating posterior parameters under local conditions. The hierarchical Bayesian approach deals with temporal changes due to government interventions (such as lockdown, school and business closures, and restricting large gatherings), changes in local conditions and operational formats (such as enabling home delivery services and working from home), and changes in individual behavior (such as social distancing, wearing a mask, and hygiene practices). It also effectively estimates the controlling parameters in the subsequent waves of a pandemic. We investigated other Bayesian network formulations based on pooling, partial pooling, and no pooling in handling the temporal variability of the pandemic. The hierarchical Bayesian model was the most effective in handling the variabilities.

The proposed model was effective in assessing risk for the COVID-19 pandemic at different geographical locations. We also analyzed the impact of distinct scenarios on the accessibility of acute care medical facilities for Ontario. We validated our findings with reported cases as well as 32 models presented by Centre for Disease Prevention and Control (CDC) for predicting the number of fatalities due to the COVID-19 in Arizona. The incorporation of the hierarchical nature

facilitate credible parameter estimation of selected periods despite the limited sample size. This work however has the following limitations.

- The present model does not explicitly consider contact tracing, vaccinated populations, super-spreading events, and voluntary preventive measures such as self-isolation, following social distancing, and wearing a mask in public places. These factors have been implicitly accounted for by using a lumped parameter approach. These factors can be included explicitly in the model.

- The model could also be improved by dividing populations based on demographics, granular subdivisions, and interaction patterns.

- In our model, we do not explicitly incorporate the inflow of additional infected people by travel. This can influence especially if the travel guidelines are not followed.

- We divided the study periods on a sequential basis without considering additional information such as holidays and super spreading events. This can be improved by dividing the periods based on the model-determined change points. The performance of the proposed hierarchical model will be compromised in case of severe changes in scenarios of the studied periods and or waves. For instance, super spreading events in a specific period could distort the model outcomes. Similarly, in studying the distinct waves using the hierarchical representation, the model would perform well if the nature (i.e, peak, growth rate) of all the subsequent waves of a location are alike.

The variability in predicting disease trajectory under evolving conditions due to distinct government regulations, societal responses, and individual practices is a major concern in pandemic risk management. The proposed hierarchical Bayesian can lead to credible estimates by effectively handling variabilities.

Acknowledgment

List of symbols and abbreviations

Symbols	Meanings
a	contagion rate
ADVI	automatic differentiation variational inference
AIC	Akaike information criterion
BBN	Bayesian belief network
BN	Bayesian network
c	infection rate
CFR	case fatality rate
COVID-19	Coronavirus disease of 2019
D, D(t)	Decesead
d_loo	relative difference between the values of LOO
DAG	directed acyclic graph
e	recovery rate
E, E(t)	Exposed
e11	fraction of the symptomatic infections during the first wave
e12	fraction of the symptomatic infections during the second wave
e21	fatality fraction of the symptomatic infections during the first wave
e22	fatality fraction of the symptomatic infections during the second wave
ELBO	evidence lower bound
f	dissimilarity function
F(X)	appearance of new infections in a compartment
HBN-SEIQRD	hierarchical Bayesian network-based susceptible, exposed, infected, quarantined, recovered, deceased model
HMC	Hamiltonian Monte Carlo
I, I(t)	Infected
LOO	leave-one-out
MCMC	Markov chain Monte Carlo
N, N(t)	total population
NELBO	negative evidence lower bound
NPIs	non-pharmaceutical interventions
NUTs	no U-turn sampler

p	real posterior distribution
p_loo	effective parameter
q	approximate the posterior
Q, Q(t)	Quarantined
R, R(t)	Recovered
R0	basic reproduction number
R0_1_2	value of the basic reproduction number between 1 and 2
R0_1_2	value of the basic reproduction number between 1 and 2
R0_1_2	value of the basic reproduction number between 1 and 2
R0_1_2_up	value of the basic reproduction number > 2
R01	reproduction number during the first wave of the outbreak
R02	reproduction number during the second wave pf the outbreak
R_{df}	disease-free reproduction number
S, S(t)	Susceptiple
SARS	severe acute respiratory syndrome
SARS-CoV-2	severe acute respiratory syndrome coronavirus-2
SEIQRD	susceptible, exposed, infected, quarantined, recovered, deceased
SEIR	susceptible, exposed, infected, and recovered
SIR	susceptible, infected, recovered
T0_3_6	value of the incubation period between 3 and 6 days
T0_7_up	value of the incubation period > 7 days
T01	incubation period in the first wave
T02	incubation period in the second wave
T1	infection period
T1_3_6	value of the infection period between 3 and 6 days
T1_6_up	value of the infection period > 6 days
T11	infection period in the first wave
T12	infection period in the second wave
T2	infection period
T3	recovery period
T3_10_14	value of the recovery period between 10 and 14 days
T3_14_up	value of the recovery period > 14 days
T31	recovery period in the first wave
T32	recovery period in the second wave
V	presence of the Coronavirus
V(X)	transfer of individuals between distinct compartments.
yij	reported data of the disease in the ith week of jth wave of the outbreak
φ_2	fraction of the quarantined/hospitalized population resulting in mortality
φ_1	fraction of symptomatic infections

θ_i	parameter of a function
$p(\theta\|y)$	probability of θ given the data
θ_{ij}	parameters of the SEIQRD model in the ith week of jth wave of the outbreak
θ	parameters of the SEIQRD model
q^*	optimal approximated posterior distribution
$KL(p\|\|q)$	KL-divergence for approximating p with q

References

Akman, D., Akman, O., & Schaefer, E. (2018). Parameter Estimation in Ordinary Differential Equations Modeling via Particle Swarm Optimization. *Journal of Applied Mathematics*, 2018, 1–9. https://doi.org/10.1155/2018/9160793

Akman, O., & Schaefer, E. (2015). An evolutionary computing approach for parameter estimation investigation of a model for cholera. *Journal of Biological Dynamics*, *9*, 147–158. https://doi.org/10.1080/17513758.2015.1039608

Alam, M., Kabir, K. M. A., & Tanimoto, J. (2020). Based on mathematical epidemiology and evolutionary game theory, which is more effective: quarantine or isolation policy. *Journal of Statistical Mechanics: Theory and Experiment*, 3, 033502. https://doi.org/10.1088/1742-5468/ab75ea

Alauddin, M., Islam Khan, M. A., Khan, F., Imtiaz, S., Ahmed, S., & Amyotte, P. (2020). How can process safety and a risk management approach guide pandemic risk management? *Journal of Loss Prevention in the Process Industries*, *68*, 104310. https://doi.org/10.1016/j.jlp.2020.104310

Alauddin, M., Khan, F., Imtiaz, S., Ahmed, S., & Amyotte, P. (2021). Pandemic risk management using engineering safety principles. *Process Safety and Environmental Protection*, *150*, 416–432. https://doi.org/10.1016/j.psep.2021.04.014

Alharthi, M., Kypraios, T., & O'Neill, P. D. (2019). Bayes factors for partially observed stochastic epidemic models. *Bayesian Analysis*, *14*(3), 907–936. https://doi.org/10.1214/18-BA1134

Allen, L. J. S. (2015). Applications of Multi-Type Branching Processes. *Stochastic Population and Epidemic Models*, 21–27. https://doi.org/10.1007/978-3-319-21554-9_3

Allen, L. J. S. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation,

and analysis. *Infectious Disease Modelling*, *2*(2), 128–142. https://doi.org/10.1016/j.idm.2017.03.001

Almutiry, W., & Deardon, R. (2020). Incorporating Contact Network Uncertainty in Individual Level Models of Infectious Disease using Approximate Bayesian Computation. *International Journal of Biostatistics*, *16*(1), 20170092. https://doi.org/10.1515/ijb-2017-0092

Amorós, R., Conesa, D., López-Quílez, A., & Martinez-Beneito, M. A. (2020). A spatio-temporal hierarchical Markov switching model for the early detection of influenza outbreaks. *Stochastic Environmental Research and Risk Assessment*, *34*(2), 275–292. https://doi.org/10.1007/s00477-020-01773-5

Anđelić, N., Šegota, S. B., Lorencin, I., Mrzljak, V., & Car, Z. (2021). Estimation of covid-19 epidemic curves using genetic programming algorithm. *Health Informatics Journal*, *27*(1), 1–40. https://doi.org/10.1177/1460458220976728

Anderson, R. M., & May, R. M. (1979). Population biology of infectious diseases: Part I. *Nature*, *280*(5721), 361–367. https://doi.org/10.1038/280361a0

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning*, *50*(1), 5–43. https://doi.org/10.1023/A:1020281327116

Apolloni, A., Poletto, C., Ramasco, J. J., Jensen, P., & Colizza, V. (2014). Metapopulation epidemic models with heterogeneous mixing and travel behaviour. *Theoretical Biology and Medical Modelling*, *11*(3). https://doi.org/10.1186/1742-4682-11-3

Arik, S.O., Li, C.-L., Yoon, J., Sinha, R., Epshteyn, A., Le, L.T., Menon, V., Singh, S., Zhang, L., Yoder, N., Nikoltchev, M., Sonthalia, Y., Nakhost, H., Kanal, E., Pfister, T., (2020). Interpretable Sequence Learning for COVID-19 Forecasting. *ArXiv*. Retrieved from http://arxiv.org/abs/2008.00646

Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., & Druzdzel, M. J. (2019). Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine. *Value in Health*, *22*(4), 439–445. https://doi.org/10.1016/j.jval.2019.01.006

Arriola, L., & Hyman, J. M. (2009). Sensitivity analysis for uncertainty quantification in mathematical models. In *Mathematical and Statistical Estimation Approaches in Epidemiology* (pp. 195–247). https://doi.org/10.1007/978-90-481-2313-1_10

Azmon, A., Faes, C., & Hens, N. (2014). On the estimation of the reproduction number based on

misreported epidemic data. *Statistics in Medicine*, *33*(7), 1176–1192. https://doi.org/10.1002/sim.6015

Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D. Y., Chen, L., & Wang, M. (2020). Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA - Journal of the American Medical Association*, *323*(14), 1406–1407. https://doi.org/10.1001/jama.2020.2565

Ball, F., & Donnelly, P. (1993). Branching Process Approximation of Epidemic Models. *Theory of Probability & Its Applications*, *37*(1), 119–121. https://doi.org/10.1137/1137024

Barrett, K., Khan, Y.A., MBiotech, S. Mac, Ximenes, R., Naimark, D.M.J., Sander, B., 2020. Estimation of covid-19induced depletion of hospital resources in ontario, Canada. *Canadian Medical Association journal,* 192(24), E640–E646. https://doi.org/10.1503/cmaj.200715

Bi et al., (2020). Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study, Lancet Infect. Dis., 20 (8), 911-919.

Birrell, P. J., de Angelis, D., & Presanis, A. M. (2018). Evidence synthesis for stochastic epidemic models. *Statistical Science*, *33*(1), 34–43. https://doi.org/10.1214/17-STS631

Bjørnstad, O. N., Finkenstädt, B. F., & Grenfell, B. T. (2002). Dynamics of measles epidemics: Estimating scaling of transmission rates using a Time series SIR model. *Ecological Monographs*, *72*(2), 169–184. https://doi.org/10.1890/0012-9615(2002)072[0169:DOMEES]2.0.CO;2

Bretó, C. (2018). Modeling and inference for infectious disease dynamics: A likelihood-based approach. *Statistical Science*, *33*(1), 57–69. https://doi.org/10.1214/17-STS636

Brown, G. D., Porter, A. T., Oleson, J. J., & Hinman, J. A. (2018). Approximate Bayesian computation for spatial SEIR(S) epidemic models. *Spatial and Spatio-Temporal Epidemiology*, *24*, 27–37. https://doi.org/10.1016/j.sste.2017.11.001

Browne, C., Gulbudak, H., & Webb, G. (2015). Modeling contact tracing in outbreaks with application to Ebola. *Journal of Theoretical Biology*, *384*, 33–49. https://doi.org/10.1016/j.jtbi.2015.08.004

Camacho, A., Kucharski, A., Aki-Sawyerr, Y., White, M.A., Flasche, S., Baguelin, M., Pollington, T., Carney, J.R., Glover, R., Smout, E., Tiffany, A., Edmunds, W.J., Funk, S. (2015). Temporal Changes in Ebola Transmission in Sierra Leone and Implications for Control Requirements: a Real-time Modelling Study. *PLoS Currents*, *Feb.*

https://doi.org/10.1371/currents.outbreaks.406ae55e83ec0b5193e30856b9235ed2

Carcione, J. M., Santos, J. E., Bagaini, C., & Ba, J. (2020). A Simulation of a COVID-19 Epidemic Based on a Deterministic SEIR Model. *Frontiers in Public Health*, *8*, 230. https://doi.org/10.3389/fpubh.2020.00230

Carley, K.M., Fridsma, D.B., Casman, E., Yahja, A., Altman, N., Chen, L.C., Kaminsky, B., Nave, D. (2006). BioWar: Scalable agent-based model of bioattacks. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, *36*(2), 252–265. https://doi.org/10.1109/TSMCA.2005.851291

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–37. https://doi.org/10.18637/jss.v076.i01

Carvalho, S. A., da Silva, S. O., & Charret, I. da C. (2019). Mathematical modeling of dengue epidemic: control methods and vaccination strategies. *Theory in Biosciences*, *138*, 223–239. https://doi.org/10.1007/s12064-019-00273-7

Cauchemez, S., Valleron, A. J., Boëlle, P. Y., Flahault, A., & Ferguson, N. M. (2008). Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, *452*, 750–754. https://doi.org/10.1038/nature06732

Chandra, V. (2020). Stochastic Compartmental Modelling of SARS-CoV-2 with Approximate Bayesian Computation. *MedRxiv*. https://doi.org/10.1101/2020.03.29.20046862

Chang, S. L., Harding, N., Zachreson, C., Cliff, O. M., & Prokopenko, M. (2020). Modelling transmission and control of the COVID-19 pandemic in Australia. *Nature communications*, *11*(1), 5710. https://doi.org/10.1038/s41467-020-19393-6

Chao, D. L., Halloran, M. E., Obenchain, V. J., & Longini, I. M. (2010). FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology*, *6*(1), e1000656. https://doi.org/10.1371/journal.pcbi.1000656

Chinazzi, M., Davis, J.T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., Pastore y Piontti, A., Mu, K., Rossi, L., Sun, K., Viboud, C., Xiong, X., Yu, H., Elizabeth Halloran, M., Longini, I.M., Vespignani, A. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, *368*(6489), 395–400. https://doi.org/10.1126/science.aba9757

Chowell, G. (2017). Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*,

2(3), 379–398. https://doi.org/10.1016/j.idm.2017.08.001

Chowell, G., & Kiskowski, M. (2016). Modeling ring-vaccination strategies to control ebola virus disease epidemics. In *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases* (pp. 71–87). https://doi.org/10.1007/978-3-319-40413-4_6

Chowell, G., Nishiura, H., & Bettencourt, L. M. A. (2007). Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface*, *4*(12), 154–166. https://doi.org/10.1098/rsif.2006.0161

Clancy, D., & Piunovskiy, A. B. (2005). An explicit optimal isolation policy for a deterministic epidemic model. *Applied Mathematics and Computation*, *163*(3), 1109–1121. https://doi.org/10.1016/j.amc.2004.06.028

Colizza, V., Barrat, A., Barthelemy, M., Valleron, A. J., & Vespignani, A. (2007). Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Medicine*, *4*(1), 0095–0110. https://doi.org/10.1371/journal.pmed.0040013

Colizza, V., & Vespignani, A. (2008). Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations. *Journal of Theoretical Biology*, *251*(3), 450–467. https://doi.org/10.1016/j.jtbi.2007.11.028

Cotta, R. M., Naveira-cotta, C. P., & Magal, P. (2020). Mathematical parameters of the COVID-19 epidemic in Brazil and evaluation of the impact of different public health measures. *Biology*, *9*(8), 220. https://doi.org/10.3390/biology9080220

Das, S. S., & Tiwari, A. K. (2021). Understanding international and domestic travel intention of Indian travellers during COVID-19 using a Bayesian approach. *Tourism Recreation Research*, *46*(2), 228–244. https://doi.org/10.1080/02508281.2020.1830341

Daszak, P. (2012). Anatomy of a pandemic. *The Lancet*, *380*(9857), 1883–1884. https://doi.org/10.1016/S0140-6736(12)61887-X

Davis, J. K., Gebrehiwot, T., Worku, M., Awoke, W., Mihretie, A., Nekorchuk, D., & Wimberly, M. C. (2019). A genetic algorithm for identifying spatially-varying environmental drivers in a malaria time series model. *Environmental Modelling and Software*, *119*, 275–284. https://doi.org/10.1016/j.envsoft.2019.06.010

de Oliveira, A. C. S., Morita, L. H. M., da Silva, E. B., Zardo, L. A. R., Fontes, C. J. F., & Granzotto, D. C. T. (2020). Bayesian modeling of COVID-19 cases with a correction to account for under-reported cases. *Infectious Disease Modelling*, *5*, 699–713.

https://doi.org/10.1016/j.idm.2020.09.005

De Visscher, A. (2020). The COVID-19 pandemic: model-based evaluation of non-pharmaceutical interventions and prognoses. *Nonlinear Dynamics*, *101*(3), 1871–1887. https://doi.org/10.1007/s11071-020-05861-7

Deardon, R., Brooks, S.P., Grenfell, B.T., Keeling, M.J., Tildesley, M.J., Savill, N.J., Shaw, D.J., Woolhouse, M.E.J. (2010). Inference for individual-level models of infectious diseases in large populations. *Statistica Sinica*, *20*(1), 239–261.

Dearlove, B., & Wilson, D. J. (2013). Coalescent inference for infectious disease: Meta-analysis of hepatitis C. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1614), 20120314. https://doi.org/10.1098/rstb.2012.0314

Dehning, J., Zierenberg, J., Spitzner, F. P., Wibral, M., Neto, J. P., Wilczek, M., & Priesemann, V. (2020). Inferring change points in the spread of COVID-19 reveals the effectiveness of interventions. *Science*, *369*(6500), eabb9789. https://doi.org/10.1126/science.abb9789

Dibble, C., & Feldmam, P. G. (2004). The GeoGraph 3D Computational Laboratory: Network and Terrain Landscapes for RePast. *Jasss*, *7*(1).

Ding, J., Hostallero, D.E., El Khili, M.R., Fonseca, G.J., Milette, S., Noorah, N., Guay-Belzile, M., Spicer, J., Daneshtalab, N., Sirois, M., Tremblay, K., Emad, A., Rousseau, S. (2021). A network-informed analysis of SARS-CoV-2 and hemophagocytic lymphohistiocytosis genes' interactions points to Neutrophil extracellular traps as mediators of thrombosis in COVID-19. *PLoS Computational Biology*, *17*(3), e1008810. https://doi.org/10.1371/journal.pcbi.1008810

Dunham, J. B. (2006). An agent-based spatially explicit epidemiological model in MASON. *Jasss*, *9*(1), 231–244.

Dureau, J., Kalogeropoulos, K., & Baguelin, M. (2013). Capturing the time-varying drivers of an epidemic using stochastic dynamical systems. *Biostatistics*, *14*(3), 541–555. https://doi.org/10.1093/biostatistics/kxs052

Eagle, N., Pentland, A. S., & Lazer, D. (2008). Mobile phone data for inferring social network structure. In *Social Computing, Behavioral Modeling, and Prediction,* pp. 79–88. https://doi.org/10.1007/978-0-387-77672-9_10

Eastman, B., Meaney, C., Przedborski, M., & Kohandel, M. (2021). Modeling the impact of public response on the COVID-19 pandemic in Ontario. *PLoS ONE*, *16*(4 April), e0249456.

https://doi.org/10.1371/journal.pone.0249456

Ejima, K, Su Kim, K., Ludema, C, Bento, A., I, Iwanami S., Fujita, Y., Ohashi, H., Koizumi, Y., Watashi, K, Aihara, K., Nishiura, H, & Iwami, S. (2021). Estimation of the incubation period of COVID-19 using viral load data, *Epidemics*, 35, 100454. https://doi.org/10.1016/j.epidem.2021.100454.

Engbert, R., Rabe, M. M., Kliegl, R., & Reich, S. (2021). Sequential Data Assimilation of the Stochastic SEIR Epidemic Model for Regional COVID-19 Dynamics. *Bulletin of Mathematical Biology*, *83*(1), 1–16. https://doi.org/10.1007/s11538-020-00834-8

Eubank, S., Guclu, H., Kumar, V. S. A., Marathe, M. V., Srinivasan, A., Toroczkai, Z., & Wang, N. (2004). Modelling disease outbreaks in realistic urban social networks. *Nature*, *429*, 180–184. https://doi.org/10.1038/nature02541

Farah, M., Birrell, P., Conti, S., & Angelis, D. De. (2014). Bayesian Emulation and Calibration of a Dynamic Epidemic Model for A/H1N1 Influenza. *Journal of the American Statistical Association*, *109*(508), 1398–1411. https://doi.org/10.1080/01621459.2014.934453

Fast, S. M., Mekaru, S., Brownstein, J. S., Postlethwaite, T. A., & Markuzon, N. (2015). The role of social mobilization in controlling ebola virus in lofa county, Liberia. *PLoS Currents*, *7*. https://doi.org/10.1371/currents.outbreaks.c3576278c66b22ab54a25e122fcdbec1

Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., Parker, M., Bonsall, D., Fraser, C. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. *Science*, *368*(6491), eabb6936. https://doi.org/10.1126/science.abb6936

Fintzi, J., Cui, X., Wakefield, J., & Minin, V. N. (2017). Efficient Data Augmentation for Fitting Stochastic Epidemic Models to Prevalence Data. *Journal of Computational and Graphical Statistics*, *26*(4), 918–929. https://doi.org/10.1080/10618600.2017.1328365

Flaxman, S., Mishra, S., Gandy, A. *et al.* Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature* **584,** 257–261 (2020). https://doi.org/10.1038/s41586-020-2405-7.

Gan, R., Tan, J., Mo, L., Li, Y., & Huang, D. (2020). Using Partial Least Squares Regression to Fit Small Data of H7N9 Incidence Based on the Baidu Index. *IEEE Access*, *8*, 60392–60400. https://doi.org/10.1109/ACCESS.2020.2983799

Gibson, G. C., Reich, N. G., & Sheldon, D. (2020). Real-time mechanistic bayesian forecasts of

COVID-19 mortality. *MedRxiv*. https://doi.org/10.1101/2020.12.22.20248736

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, M. (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, *26*(6), 855–860. https://doi.org/10.1038/s41591-020-0883-7

He, S., Peng, Y., & Sun, K. (2020). SEIR modeling of the COVID-19 and its dynamics. *Nonlinear Dynamics*, *101*(3), 1667–1680. https://doi.org/10.1007/s11071-020-05743-y

Hethcote, H. W. (1976). Qualitative analyses of communicable disease models. *Mathematical Biosciences*, *28*(3–4), 335–356. https://doi.org/10.1016/0025-5564(76)90132-2

Hiorns, R. W., & MacDonald, N. (1982). Time Lags in Biological Models. *Journal of the Royal Statistical Society. Series A (General)*, *145*(1), 140. https://doi.org/10.2307/2981435

Hoffman, M. D., & Gelman, A. (2014). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hossain, M. P., Junus, A., Zhu, X., Jia, P., Wen, T. H., Pfeiffer, D., & Yuan, H. Y. (2020). The effects of border control and quarantine measures on the spread of COVID-19. *Epidemics*, 32, 100397. https://doi.org/10.1016/j.epidem.2020.100397

Hosseini, E., Ghafoor, K. Z., Sadiq, A. S., Guizani, M., & Emrouznejad, A. (2020). COVID-19 Optimizer Algorithm, Modeling and Controlling of Coronavirus Distribution Process. *IEEE Journal of Biomedical and Health Informatics*, *24*(10), 2765–2775. https://doi.org/10.1109/JBHI.2020.3012487

House, T., & Keeling, M. J. (2008). Deterministic epidemic models with explicit household structure. *Mathematical Biosciences*, *213*(1), 29–39. https://doi.org/10.1016/j.mbs.2008.01.011

Hu, Z., Cui, Q., Han, J., Wang, X., Sha, W. E. I., & Teng, Z. (2020). Evaluation and prediction of the COVID-19 variations at different input population and quarantine strategies, a case study in Guangdong province, China. *International Journal of Infectious Diseases*, *95*, 231–240. https://doi.org/10.1016/j.ijid.2020.04.010

Isa Irawan, M., & Amiroch, S. (2015). Construction of phylogenetic tree using neighbor joining algorithms to identify the host and the spreading of SARS epidemic. *Journal of Theoretical and Applied Information Technology*, *71*(3), 424–429.

Ivorra, B., Ferrández, M. R., Vela-Pérez, M., & Ramos, A. M. (2020). Mathematical modeling of

the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China. *Communications in Nonlinear Science and Numerical Simulation*, *88*, 105303. https://doi.org/10.1016/j.cnsns.2020.105303

Jenness, S. M., Goodreau, S. M., & Morris, M. (2018). Epimodel: An R package for mathematical modeling of infectious disease over networks. *Journal of Statistical Software*, *84*(8). https://doi.org/10.18637/jss.v084.i08

Jung, E., Lee, J., & Chowell, G. (2016). Evaluating the number of sickbeds during ebola epidemics using optimal control theory. In *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases* (pp. 89–101). https://doi.org/10.1007/978-3-319-40413-4_7

Kao, Y. H., & Eisenberg, M. C. (2018). Practical unidentifiability of a simple vector-borne disease model: Implications for parameter estimation and intervention assessment. *Epidemics*, *25*, 89–100. https://doi.org/10.1016/j.epidem.2018.05.010

Kermack, W.O., McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *115*(772), 700–721. https://doi.org/10.1098/rspa.1927.0118

Khan, A., Hussain, G., Zahri, M., Zaman, G., & Wannasingha Humphries, U. (2020). A stochastic SACR epidemic model for HBV transmission. *Journal of Biological Dynamics*, *14*(1), 788–801. https://doi.org/10.1080/17513758.2020.1833993

Kim, S., Ko, Y., Kim, Y. J., & Jung, E. (2020). The impact of social distancing and public behavior changes on COVID-19 transmission dynamics in the Republic of Korea. *PLoS ONE*, *15*(9 September 2020), e0238684. https://doi.org/10.1371/journal.pone.0238684

Kucukelbir, A., Blei, D. M., Gelman, A., Ranganath, R., & Tran, D. (2017). Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, *18*, 1–45.

Kumar, A., Goel, K., & Nilam. (2020). A deterministic time-delayed SIR epidemic model: mathematical modeling and analysis. *Theory in Biosciences*, *139*(1), 67–76. https://doi.org/10.1007/s12064-019-00300-7

Kypraios, T., Neal, P., & Prangle, D. (2017). A tutorial introduction to Bayesian inference for stochastic epidemic models using Approximate Bayesian Computation. *Mathematical Biosciences*, *287*, 42–53. https://doi.org/10.1016/j.mbs.2016.07.001

Lange, A. (2016). Reconstruction of disease transmission rates: Applications to measles, dengue, and influenza. *Journal of Theoretical Biology*, *400*, 138–153.

https://doi.org/10.1016/j.jtbi.2016.04.017

Lanzas, C., & Chen, S. (2015). Complex system modelling for veterinary epidemiology. *Preventive Veterinary Medicine*, *118*(2–3), 207–214. https://doi.org/10.1016/j.prevetmed.2014.09.012

Lawson, A. B. (2020). NIMBLE for Bayesian Disease Mapping. *Spatial and Spatio-Temporal Epidemiology*, *33*, 100323. https://doi.org/10.1016/j.sste.2020.100323

Lee, S. Y., Lei, B., & Mallick, B. (2020). Estimation of COVID-19 spread curves integrating global data and borrowing information. *PLoS ONE*, *15*(7 July), 1–17. https://doi.org/10.1371/journal.pone.0236860

Lee, V. J., Lye, D. C., & Wilder-Smith, A. (2009). Combination strategies for pandemic influenza response - a systematic review of mathematical modeling studies. *BMC Medicine*, *7,* 76. https://doi.org/10.1186/1741-7015-7-76

Legrand, J., Grais, R. F., Boelle, P. Y., Valleron, A. J., & Flahault, A. (2007). Understanding the dynamics of Ebola epidemics. *Epidemiology and Infection*, *135*(4), 610–621. https://doi.org/10.1017/S0950268806007217

Lekone, P. E., & Finkenstädt, B. F. (2006). Statistical inference in a stochastic epidemic SEIR model with control intervention: Ebola as a case study. *Biometrics*, *62*(4), 1170–1177. https://doi.org/10.1111/j.1541-0420.2006.00609.x

Li, M., Dushoff, J., & Bolker, B. M. (2018). Fitting mechanistic epidemic models to data: A comparison of simple Markov chain Monte Carlo approaches. *Statistical Methods in Medical Research*, *27*(7), 1956–1967. https://doi.org/10.1177/0962280217747054

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S.M., Lau, E.H.Y., Wong, J.Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T.T.Y., Wu, J.T., Gao, G.F., Cowling, B.J., Yang, B., Leung, G.M., Feng, Z. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. *New England Journal of Medicine*, *382*(13), 1199–1207. https://doi.org/10.1056/nejmoa2001316

Lim, J., Akashi, Y., Song, D., Hwang, H., Kuwahara, Y., Yamamura, S., Yoshimoto, N., Itahashi, K. (2018). Hierarchical Bayesian modeling for predicting ordinal responses of personalized thermal sensation: Application to outdoor thermal sensation data. *Building and Environment*,

142, 414–426. https://doi.org/10.1016/j.buildenv.2018.06.045

Lin, Q., Zhao, S., Gao, D., Lou, Y., Yang, S., Musa, S.S., Wang, M.H., Cai, Y., Wang, W., Yang, L., He, D. (2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action. *International Journal of Infectious Diseases*, *93*, 211–216. https://doi.org/10.1016/j.ijid.2020.02.058

Lipsitch, M., Cohen, T., Cooper, B., Robins, J.M., Ma, S., James, L., Gopalakrishna, G., Chew, S.K., Tan, C.C., Samore, M.H., Fisman, D., Murray, M. (2003). Transmission dynamics and control of severe acute respiratory syndrome. *Science*, *300*(5627), 1966–1970. https://doi.org/10.1126/science.1086616

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, *27*(2), 1–4. https://doi.org/10.1093/jtm/taaa021

Lytras, T., Gkolfinopoulou, K., Bonovas, S., & Nunes, B. (2019). FluHMM: A simple and flexible Bayesian algorithm for sentinel influenza surveillance and outbreak detection. *Statistical Methods in Medical Research*, *28*(6), 1826–1840. https://doi.org/10.1177/0962280218776685

Mallapaty, S. (2021). What the data say about border closures and COVID spread. *Nature,* 589, 185. https://doi.org/10.1038/d41586-020-03605-6

Mallick, M. R., & Imtiaz, S. A. (2013). A hybrid method for process fault detection and diagnosis. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 46*(32),* 827–832. https://doi.org/10.3182/20131218-3-IN-2045.00099

Manevski, D., Ružić Gorenjec, N., Kejžar, N., & Blagus, R. (2020). Modeling COVID-19 pandemic using Bayesian analysis with application to Slovene data. *Mathematical Biosciences*, 329, 108466. https://doi.org/10.1016/j.mbs.2020.108466

Martínez-Beneito, M. A., Conesa, D., López-Quílez, A., & López-Maside, A. (2008). Bayesian Markov switching models for the early detection of influenza epidemics. *Statistics in Medicine*, *27*(22), 4455–4468. https://doi.org/10.1002/sim.3320

Martínez-López, B., Perez, A. M., & Sánchez-Vizcaíno, J. M. (2009). Social network analysis. Review of general concepts and use in preventive veterinary medicine. *Transboundary and Emerging Diseases*, *56*(4), 109–120. https://doi.org/10.1111/j.1865-1682.2009.01073.x

Marwa, Y. M., Mwalili, S., & Mbalawata, I. S. (2019). Bayesian approach in modelling cholera

outbreak in Ilala municipal council, Tanzania. *Journal of Mathematical and Computational Science*, *9*(1), 73–86. https://doi.org/10.28919/jmcs/3881

McKinley, T.J., Vernon, I., Andrianakis, I., McCreesh, N., Oakley, J.E., Nsubuga, R.N., Goldstein, M., White, R.G. (2018). Approximate bayesian computation and simulation-based inference for complex stochastic epidemic models. *Statistical Science*, *33*(1), 4–18. https://doi.org/10.1214/17-STS618

Minter, A., & Retkute, R. (2019). Approximate Bayesian Computation for infectious disease modelling. *Epidemics*, *29*(August), 100368. https://doi.org/10.1016/j.epidem.2019.100368

Mniszewski, S. M., Del Valle, S. Y., Stroud, P. D., Riese, J. M., & Sydoriak, S. J. (2008). EpiSimS simulation of a multi-component strategy for pandemic influenza. In *Proceedings of the 2008 Spring Simulation Multiconference, SpringSim'08* (pp. 556–563). https://doi.org/10.1145/1400549.1400636

Monteiro, L. H. A., Gandini, D. M., & Schimit, P. H. T. (2020). The influence of immune individuals in disease spread evaluated by cellular automaton and genetic algorithm. *Computer Methods and Programs in Biomedicine*, *196*, 105707. https://doi.org/10.1016/j.cmpb.2020.105707

Neal, P. J. (2019). Approximate Bayesian Computation Methods for Epidemic Models. In *Handbook of Infectious Disease Data Analysis* (pp. 179–197). https://doi.org/10.1201/9781315222912-10

Neal, P., & Terry Huang, C. L. (2015). Forward Simulation Markov Chain Monte Carlo with Applications to Stochastic Epidemic Models. *Scandinavian Journal of Statistics*, *42*(2), 378–396. https://doi.org/10.1111/sjos.12111

Nishio, M., & Arakawa, A. (2019). Performance of Hamiltonian Monte Carlo and No-U-Turn Sampler for estimating genetic parameters and breeding values. *Genetics Selection Evolution*, *51*, 73. https://doi.org/10.1186/s12711-019-0515-1

Nissen, K., Hagbom, M., Krambrich, J., Akaberi, D., Sharma, S., Ling, J., Hoffman, T., Svensson, L., Bondeson, K., Salaneck, E. (2021). Presymptomatic viral shedding and infective ability of SARS-CoV-2; a case report. *Heliyon*, *7*(2), e06328. https://doi.org/10.1016/j.heliyon.2021.e06328

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences*, *180*(1–2), 103–114.

https://doi.org/10.1016/S0025-5564(02)00109-8

Okyere, E., De-Graft Ankamah, J., Hunkpe, A. K., & Mensah, D. (2020). Deterministic Epidemic Models for Ebola Infection with Time-Dependent Controls. *Discrete Dynamics in Nature and Society*, *2020*, 1–12. https://doi.org/10.1155/2020/2823816

Osei, F. B., Duker, A. A., & Stein, A. (2012). Bayesian structured additive regression modeling of epidemic data: Application to cholera. *BMC Medical Research Methodology*, *12, 118*. https://doi.org/10.1186/1471-2288-12-118

Otoo, D., Opoku, P., Charles, S., & Kingsley, A. P. (2020). Deterministic epidemic model for (SVCSyCAsyIR) pneumonia dynamics, with vaccination and temporal immunity. *Infectious Disease Modelling*, *5*, 42–60. https://doi.org/10.1016/j.idm.2019.11.001

Paiva, H. M., Afonso, R. J. M., de Oliveira, I. L., & Garcia, G. F. (2020). A data-driven model to describe and forecast the dynamics of COVID-19 transmission. *PLoS ONE*, *15*(7), e0236386. https://doi.org/10.1371/journal.pone.0236386

Pedroso-Rodriguez, L. M., Marrero, A., & De Arazoza, H. (2003). Nonlinear parametric model identification using genetic algorithms. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2687*, 473–480. https://doi.org/10.1007/3-540-44869-1_60

Pell, B., Baez, J., Phan, T., Gao, D., Chowell, G., & Kuang, Y. (2016). Patch models of EVD transmission dynamics. In *Mathematical and Statistical Modeling for Emerging and Re-emerging Infectious Diseases* (pp. 147–167). https://doi.org/10.1007/978-3-319-40413-4_10

Plummer, M. (2003). JAGS : A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing; Vienna, Austria*, (Dsc).

Pollicott, M., Wang, H., & Weiss, H. (2012). Extracting the time-dependent transmission rate from infection data via solution of an inverse ODE problem. *Journal of Biological Dynamics*, *6*(2), 509–523. https://doi.org/10.1080/17513758.2011.645510

Polo, G., Mera Acosta, C., Soler-Tovar, D., Porras Villamil, J.F., Palencia, N.P., Penagos, M., Meza Martinez, J., Bobadilla, J.N., Martin, L.V., Durán, S., Rodriguez Álvarez, M., Meza Carvajalino, C., Villamil, L.C., Benavides Ortiz, E. (2020). Bayesian Spatio-Temporal Modeling of COVID-19: Inequalities on Case-Fatality Risk. *MedRxiv*. https://doi.org/10.1101/2020.08.18.20171074

Ponciano, J. M., & Capistrán, M. A. (2011). First principles modeling of nonlinear incidence rates in seasonal epidemics. *PLoS Computational Biology*, *7*(2), e1001079. https://doi.org/10.1371/journal.pcbi.1001079

Popinga, A., Vaughan, T., Stadler, T., & Drummond, A. J. (2014). Inferring epidemiological dynamics with Bayesian coalescent inference: The merits of deterministic and stochastic models. *Genetics*, *199*(2), 595–607. https://doi.org/10.1534/genetics.114.172791

Price, D. J., Bean, N. G., Ross, J. V., & Tuke, J. (2016). On the efficient determination of optimal Bayesian experimental designs using ABC: A case study in optimal observation of epidemics. *Journal of Statistical Planning and Inference*, *172*, 1–15. https://doi.org/10.1016/j.jspi.2015.12.008

Rachah, A., & Torres, D. F. M. (2016). Dynamics and Optimal Control of Ebola Transmission. *Mathematics in Computer Science*, *10*(3), 331–342. https://doi.org/10.1007/s11786-016-0268-y

Rivers, C. M., Lofgren, E. T., Marathe, M., Eubank, S., & Lewis, B. L. (2014). Modeling the Impact of Interventions on an Epidemic of Ebola in Sierra Leone and Liberia. *PLoS Currents*, 6. https://doi.org/10.1371/currents.outbreaks.4d41fe5d6c05e9df30ddce33c66d084c

Rizzo, C., & Atti, M. L. C. degli. (2008). Modeling influenza pandemic and interventions. In *Influenza Vaccines for the Future* (pp. 281–296). https://doi.org/10.1007/978-3-7643-8371-8_13

Romero-Severson, E. O., Ribeiro, R. M., & Castro, M. (2018). Noise is not error: Detecting parametric heterogeneity between epidemiologic time series. *Frontiers in Microbiology*, *9*, 1529. https://doi.org/10.3389/fmicb.2018.01529

Rozell, D. J. (2019). A Critique of Pandemic Catastrophe Modeling. *Journal of Risk Analysis and Crisis Response*, *9*(3), 128-133. https://doi.org/10.2991/jracr.k.191024.002

Russell, T.W., Golding, N., Hellewell, J., Abbott, S., Wright, L., Pearson, C.A.B., van Zandvoort, K., Jarvis, C.I., Gibbs, H., Liu, Y., Eggo, R.M., Edmunds, W.J., Kucharski, A.J., Deol, A.K., Villabona-Arenas, C.J., Jombart, T., O'Reilly, K., Munday, J.D., Meakin, S.R., Lowe, R., Gimma, A., Endo, A., Nightingale, E.S., Medley, G., Foss, A.M., Knight, G.M., Prem, K., Hué, S., Diamond, C., Rudge, J.W., Atkins, K.E., Auzenbergs, M., Flasche, S., Houben, R.M.G.J., Quilty, B.J., Klepac, P., Quaife, M., Funk, S., Leclerc, Q.J., Emery, J.C., Jit, M., Simons, D., Bosse, N.I., Procter, S.R., Sun, F.Y., Clifford, S., Sherratt, K., Rosello, A.,

Davies, N.G., Brady, O., Tully, D.C., Gore-Langton, G.R. (2020). Reconstructing the early global dynamics of under-ascertained COVID-19 cases and infections. *BMC Medicine*, *18*(1), 332. https://doi.org/10.1186/s12916-020-01790-9

Santermans, E., Robesyn, E., Ganyani, T., Sudre, B., Faes, C., Quinten, C., Van Bortel, W., Haber, T., Kovac, T., Van Reeth, F., Testa, M., Hens, N., Plachouras, D. (2016). Spatiotemporal evolution of Ebola virus disease at sub-national level during the 2014 West Africa epidemic: Model scrutiny and data meagreness. *PLoS ONE*, *11*(1), e0147172. https://doi.org/10.1371/journal.pone.0147172

Shangguan, D., Liu, Z., Wang, L., & Tan, R. (2021). A stochastic epidemic model with infectivity in incubation period and homestead–isolation on the susceptible. *Journal of Applied Mathematics and Computing*, 15, 1–21. https://doi.org/10.1007/s12190-021-01504-1

Sharkey, K. J. (2011). Deterministic epidemic models on contact networks: Correlations and unbiological terms. *Theoretical Population Biology*, *79*(4), 115–129. https://doi.org/10.1016/j.tpb.2011.01.004

Sharmin, S., Glass, K., Viennet, E., & Harley, D. (2018). A Bayesian approach for estimating under-reported dengue incidence with a focus on non-linear associations between climate and dengue in Dhaka, Bangladesh. *Statistical Methods in Medical Research*, *27*(4), 991–1000. https://doi.org/10.1177/0962280216649216

Shen, M., Xiao, Y., & Rong, L. (2015). Modeling the effect of comprehensive interventions on Ebola virus transmission. *Scientific Reports*, *5,* 15818. https://doi.org/10.1038/srep15818

Short, K. R., Kedzierska, K., & van de Sandt, C. E. (2018). Back to the Future: Lessons Learned From the 1918 Influenza Pandemic. *Frontiers in Cellular and Infection Microbiology*, *8*, 343. https://doi.org/10.3389/fcimb.2018.00343

Smirnova, A., deCamp, L., & Chowell, G. (2019). Forecasting Epidemics Through Nonparametric Estimation of Time-Dependent Transmission Rates Using the SEIR Model. *Bulletin of Mathematical Biology*, *81*(11), 4343–4365. https://doi.org/10.1007/s11538-017-0284-3

Subramanian, R., He, Q., & Pascual, M. (2021). Quantifying asymptomatic infection and transmission of COVID-19 in New York City using observed cases, serology, and testing capacity. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(9), e2019716118. https://doi.org/10.1073/pnas.2019716118

Taghizadeh, L., Karimi, A., & Heitzinger, C. (2020). Uncertainty quantification in epidemiological

models for the COVID-19 pandemic. *Computers in Biology and Medicine*, *125,* 104011. https://doi.org/10.1016/j.compbiomed.2020.104011

Taylor, B. P., Dushoff, J., & Weitz, J. S. (2016). Stochasticity and the limits to confidence when estimating R0 of Ebola and other emerging infectious diseases. *Journal of Theoretical Biology*, *408*, 145–154. https://doi.org/10.1016/j.jtbi.2016.08.016

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234-240. https://doi.org/10.2307/143141

Tong, Z.D., Tang, A., Li, K.F., Li, P., Wang, H.L., Yi, J.P., Zhang, Y.L., Yan, J.B. (2020). Potential presymptomatic transmission of SARS-CoV-2, Zhejiang Province, China, 2020. *Emerging Infectious Diseases*, *26*(5), 1052–1054. https://doi.org/10.3201/eid2605.200198

Touloupou, P., Finkenstädt, B., & Spencer, S. E. F. (2020). Scalable Bayesian Inference for Coupled Hidden Markov and Semi-Markov Models. *Journal of Computational and Graphical Statistics*, *29*(2), 238–249. https://doi.org/10.1080/10618600.2019.1654880

Trawicki, M. B. (2017). Deterministic seirs epidemic model for modeling vital dynamics, vaccinations, and temporary immunity. *Mathematics*, *5*(1), 1–7. https://doi.org/10.3390/math5010007

Tritch, W., & Allen, L. J. S. (2018). Duration of a minor epidemic. *Infectious Disease Modelling*, *3*, 60–73. https://doi.org/10.1016/j.idm.2018.03.002

Turkkan, N., and T. Pham-Gia. "Algorithm AS 308: Highest posterior density credible region and minimum area confidence region: the Bivariate case." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 46(1), 1997, 131–40. http://www.jstor.org/stable/2986210.

Vazquez-Prokopec, G.M., Bisanzio, D., Stoddard, S.T., Paz-Soldan, V., Morrison, A.C., Elder, J.P., Ramirez-Paredes, J., Halsey, E.S., Kochel, T.J., Scott, T.W., Kitron, U. (2013). Using GPS Technology to Quantify Human Mobility, Dynamic Contacts and Infectious Disease Dynamics in a Resource-Poor Urban Environment. *PLoS ONE*, *8*(4), e58802. https://doi.org/10.1371/journal.pone.0058802

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Venkatramanan, S., Lewis, B., Chen, J., Higdon, D., Vullikanti, A., & Marathe, M. (2018). Using data-driven agent-based models for forecasting emerging infectious diseases. *Epidemics*, *22*,

43–49. https://doi.org/10.1016/j.epidem.2017.02.010

Wang, W., Ji, C., Bi, Y., & Liu, S. (2020). Stability and asymptoticity of stochastic epidemic model with interim immune class and independent perturbations. *Applied Mathematics Letters*, *104*, 106245. https://doi.org/10.1016/j.aml.2020.106245

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Wei, W. E., Li, Z., Chiew, C. J., Yong, S. E., Toh, M. P., & Lee, V. J. (2020). Presymptomatic Transmission of SARS-CoV-2 — Singapore, January 23–March 16, 2020. *MMWR. Morbidity and Mortality Weekly Report*, *69*(14), 411–415. https://doi.org/10.15585/mmwr.mm6914e1

White, L. F., & Pagano, M. (2008). A likelihood-based method for real-time estimation of the serial interval and reproductive number of an epidemic. *Statistics in Medicine*, *27*(16), 2999–3016. https://doi.org/10.1002/sim.3136

Wu, K. M., & Riley, S. (2016). Estimation of the basic reproductive number and mean serial interval of a novel pathogen in a small, well-observed discrete population. *PLoS ONE*, *11*(2), e0148061. https://doi.org/10.1371/journal.pone.0148061

Xu, X., Kypraios, T., & O'Neill, P. D. (2016). Bayesian non-parametric inference for stochastic epidemic models using Gaussian Processes. *Biostatistics*, *17*(4), 619–633. https://doi.org/10.1093/biostatistics/kxw011

Yan, X., & Zou, Y. (2008). Optimal and sub-optimal quarantine and isolation control in SARS epidemics. *Mathematical and Computer Modelling*, *47*(1), 235–245. https://doi.org/10.1016/j.mcm.2007.04.003

Yang, H. M. (2014). The basic reproduction number obtained from Jacobian and next generation matrices - A case study of dengue transmission modelling. *BioSystems*, *126*, 52–75. https://doi.org/10.1016/j.biosystems.2014.10.002

Zhou, L., Miranda-Saksena, M., & Saksena, N. K. (2013). Viruses and neurodegeneration. *Virology Journal*, *10*(1), 172. https://doi.org/10.1186/1743-422X-10-172

Zhou, X., Li, X., & Wang, W. S. (2014). Bifurcations for a deterministic SIR epidemic model in discrete time. *Advances in Difference Equations*, *2014*(1), 168. https://doi.org/10.1186/1687-1847-2014-168.

Chapter 8

Pandemic Risk Management using Engineering Safety Principles

Preface: This chapter is furthering our hypothesis of improving data-driven methods using extensive knowledge and science-based formulations. Integration based on Engineering approaches such as precautionary and ALARP principles along with the mechanistic model of the disease dynamics has been used to assess risk under distinct scenarios. Uncertainties in the parameters have been captured using Monte Carlo Simulation. This work can be mapped to the sub-objective of the thesis "to establish synergy between process safety and pandemic risk management". The chapter has been published in Process Safety and Environmental Protection, 150, 416-432.

Abstract: The containment of infectious diseases is challenging due to complex transmutation in the biological system, intricate global interactions, intense mobility, and multiple transmission modes. An emergent disease has the potential to turn into a pandemic impacting millions of people with loss of life, mental health, and severe economic impairment. Multifarious approaches to risk management have been explored for combating an epidemic spread. This work presents engineering safety principles to pandemic risk management. We have assessed the pandemic risk using Paté-Cornell's six levels of uncertainty. The susceptible, exposed, infected, quarantined, recovered, deceased (SEIQRD) along with the Monte Carlo simulation has been used to estimate the associated risk of a pandemic. The risk minimization strategies have been categorized into hierarchical safety measures. We have developed an event tree model of pandemic risk management for distinct risk-reducing strategies realized due to natural evolution, government interventions, societal responses, and individual practices. The roles of distinct interventions have also been investigated for the survival of infected individuals under the existing healthcare facilities. We have studied the Coronavirus disease of 2019 (COVID-19) for pandemic risk management using the proposed framework. The results highlight effectiveness of the proposed strategies in containing a pandemic.

Keywords: Risk analysis, pandemic, non-pharmaceutical interventions, precautionary principle, ALARP, COVID-19.

8.1 Introduction

The global pandemic of coronavirus disease (COVID-19) is affecting billions of people worldwide with public health, livelihood, food security, fear, and suffering. Mortality, compromised mental health, and employment loss are its immediate impacts; the pandemic's long-term repercussions will be a crisis in public finance, including debt and fiscal rebalancing. The COVID-19 pandemic has caused more than 120 million infected cases and over 2.5 million mortalities to date (Worldometer, March 14, 2021). The World Bank's economic forecast indicates that the pandemic could dramatically reduce the gross domestic product (GDP) worldwide (World Bank, 2020). The

COVID-19's social and economic disruption is devastating; almost half of the global workforce is at risk of loss of livelihoods, tens of millions of people are in danger of falling into extreme poverty, and millions of enterprises are facing an existential threat (Joint statement by ILO, FAP, and WHO, October 13, 2020).

Vaccination is a proven method for adequate protection: however, the development, production, and distribution of a vaccine requires several months. For instance, the dosage administered to date (March, 2021) for the COVID-19 pandemic can meet only 3.1% of the global population (Bloomberg, 2021). Many non-pharmaceutical interventions (NPIs) have been effective in controlling the spread of a pandemic to an acceptable level. Isolation, social distancing, putting on personal protective equipment (PPE), and following good hygiene practices, e.g., frequent hand washing and refraining from face touching, are key non-pharmaceutical strategies for containing the epidemic spread (Ferguson et al., 2020; Davies et al., 2020). Government interventions such as lockdown, school and business closures, and a ban on social gatherings are other effective measures for containing the disease spread. The early detection of the infected cases, contact tracing, and quarantine of exposed cases are effective strategies for restricting the spread of a pandemic. The time frame of implementing and relaxing interventions also plays a vital role in controlling the epidemic (Alauddin et al., 2020). However, these preventive measures have unwanted socio-economic consequences including loss of income, poor mental health, and domestic violence. Therefore, there is a crucial need for balancing risk and benefits.

 Risk assessment is crucial in many disciplines, e.g., engineering and infrastructure, exposure assessment, process safety, occupational health and safety, risk policy and legislation, and security and defense (Aven, 2016). Risk assessment guides decision-making in selecting alternatives, approving practices, and implementing risk-reducing measures. Several risk

analyses techniques, e.g. failure mode and effects analysis (FMEA), hazard and operability study (HAZOP), fault tree analysis (FTA), event tree analysis (ETA), bow-tie analysis (BTA), Markov chain analysis (MCA), and Bayesian networks (BNs) have been used for assessing risk of processing systems (Cameron et al., 2017; Cui, Zhao, & Zhang, 2010; Khakzad, Khan, & Amyotte, 2013; Khan, Rathnayaka, & Ahmed, 2015; Zhang, Wu, Hu, & Ni, 2018). FTA and ETA are two well-established risk assessment methods for providing qualitative analysis of hazards identification and quantitative assessment of likelihood. "Bow-tie" combines the FTA and ETA by a common top-event named as a critical event (Khakzad et al., 2013; Xin, Zhang, Jin, & Zhang, 2019). The layer of protection analysis (LOPA) and inherently safer design (ISD) are the other promising risk assessment and management tools. Public awareness profoundly affects public policy development for risk management (Pike, Khan, & Amyotte, 2020). Renn (1998) proposed a public participation model based on integrating analytic knowledge and risk perception. Decision analysis tools such as cost-benefit analysis, cost-effectiveness analysis, and multi-attribute analysis help evaluate relative risk in the risk assessment (Aven, 2016; Kabyl, Yang, Abbassi, & Li, 2020).

Uncertainty is critical in risk conceptualization and risk assessments. Uncertainties can be categorized as aleatory (come from the variability in population/ data) and epistemic (arises from lack of knowledge of the phenomena) (He et al., 2018). Paté-Cornell (1996) proposed six treatment levels of both aleatory and epistemic types of uncertainty for risk analyses. Spiegelhalter and Riesch (2011) categorized uncertainty into five levels: event, parameter, model, acknowledged, and unknown inadequacies. The adaptive risk management approach to estimate high uncertainties was conferred by Cox (2012). The elements of the high and low levels of uncertainty have been displayed in Table 8.1 (Goerlandt & Reniers, 2016).

Table 8.1: The constituents of a high and low level of uncertainty

Low Uncertainty	High Uncertainty
1. Highly reasonable assumptions	1. Strong and overly simplified assumptions
2. Reliable data	2. Unreliable data
3. Consensus among experts	3. Lack of consensus among experts
4. Well understood phenomena	4. Obscure phenomena

To deal with uncertainties, the cautionary/precautionary techniques, also referred to as strategies of robustness, have been universally applied for minimizing risk in many disciplines (Aven, 2016). These principles are based on the development of substitutes, redundancy in designing safety devices, and safety factors. The ALARP (as low as reasonably practicable) principle is a risk-reduction principle based on both risk-informed and cautionary/precautionary thinking. The ALARP principle is a fundamental approach to assessing tolerable risk. The approach sets an upper limit above which the risk must be reduced or the activity must stop and a lower limit below which resources expended bring negligible risk reduction (Pike, Khan, & Amyotte, 2020).

The dynamic behavior of a process system and an epidemic has many similarities. Compartmental models have been employed to model the dynamics of many chemical processing systems, e.g., the Fischer-Tropsch synthesis (FTS) (Iliuta et al, 2007), bioprocess design (Cui et al 1996, Laakkonen et al, 2006, and Vrábel et al, 1999), crystallization (Bermingham, Kramer, & Rosmalen, 1998; Irizarry-Revera, 2012), precipitation (Zhao et al., 2017) as well as waste treatment (Alvarado, Vedantam, Goethals, & Nopens, 2012). Alauddin et al. (2020) presented similarities between the SIR epidemiological model and the reaction kinetics model of a CSTR by highlighting resemblance in the conservation principles and factors governing the contagion and reaction rates.

The methodologies to prevent, control, and mitigate infection are analogous to the hazard control and safety frameworks used in the process industries. Different safety barriers such as basic process control, alarms and operator interventions, safety instrumented systems, relief devices, and physical containments are used as control layers for the abnormal situation management of chemical processes (Dowell, 1999; Willey, 2014). Brown, Amyotte and Vanberkel (2021) classified distinct measures of restraining epidemic diseases into hierarchical process safety principles. Lindhout and Reniers (2020) proposed an integrated pandemics barrier model based on sequential steps of an outbreak. They described what could have been done better in preventing and repressing the Covid-19 pandemic from a safety management perspective. Alauddin et al (2020) developed a layer of protection analysis (LOPA) for preventive, controlling, and mitigating strategies for pandemic risk. Also, several areas of similarities were identified where process safety and epidemiology could benefit from each other. These include (i) early fault detection vs early case detection, (ii) identification of effective control mechanism, (iii) the fast response of public health vs operator response, (iv) effective resource allocation and mobilization, (v) identification of the most vulnerable elements, and (vi) application of expertise from similar outbreaks in the past vs use of historical process data.

Engineering safety protocols are applicable to pandemic risk management to a great extent. The present pandemic also offers many learning opportunities to improve engineering risk management practices. By drawing a parallel between the two domains, we believe that the lessons learned from the COVID-19 pandemic would immensely benefit engineering safety personnel and healthcare experts in efficient policymaking.

The objective of this work is to apply some of the techniques used in process safety analysis and risk mitigation in pandemic risk management. Specifically, we have focused on applying the

precautionary and ALARP approaches for evaluating the risk of infectious diseases. The contribution of this paper includes:

i. *Pandemic risk analysis using the precautionary principle:* We have analyzed the risk of COVID-19 using Paté-Cornell's six levels of qualitative and quantitative analysis. The SEIQRD pandemic model and the Monte Carlo simulation have been used for risk estimation.

ii. *Development of event tree diagram for pandemic risk management:* Many Risk-reducing strategies realized due to natural evolution, government interventions, and individual practices have been presented as barriers to minimize the pandemic risk.

iii. *Risk analysis using ALARP:* The enforcement of risk-reducing measures, including no measure, has been studied using the ALARP approach to risk assessment. We have assessed the quantitative risk estimated using the SEIQRD model. The uncertainty in the parameters has been accounted for by the Monte Carlo simulation.

iv. *Survival analysis in COVID-19 under the existing healthcare systems:* The existing healthcare system's sufficiency depends upon the effectiveness of the strategies for restraining a pandemic. The survival analysis of an infected individual due to the availability of treatment under the current healthcare system has been analyzed under different strategies.

Section 8.2 provides the mathematical model for the epidemic spread, including the risk management approaches. We have presented the SEIQRD model, followed by a brief discussion on the precautionary and the ALARP approaches. The risk assessment of COVID-19 for distinct scenarios is presented in Section 8.3. Finally, Section 8.4 concludes with findings, limitations, and scopes for future work.

8.2 Methods and models

8.2.1: The SEIQRD model

Compartmental models have been widely used for the prediction and control of pandemics. They are based on systems of ordinary differential equations that focus on the dynamic progression of a population through different epidemiological states (Chowell, 2017). The population is divided into distinct compartments, each having the same state of the epidemic. The SIR (susceptible, infected, recovered) model suggests that the infected hosts become contagious immediately after exposure to an infected carrier (Anderson & May, 1979; Hethcote, 1976; Hiorns & MacDonald, 1982; Kermack & McKendrick, 1927). The latency period, the period between exposures and infectious, is taken into account by SEIR (susceptible, exposed, infected, recovered) model. Many extended compartment models have been developed to take into account isolation, quarantine, and hospitalization (Alauddin et al., 2020; Giordano et al., 2020; Hu et al., 2020; Li et al., 2020; Lin et al., 2020; Paiva, Afonso, de Oliveira, & Garcia, 2020). The SEIQRD model captures the effect of hospitalization and quarantine on the disease spread (Fig. 8.1). The mathematical formulations of the SEIQRD model are presented in Eqs. 8.1-8.7, where 'a', 'b', 'c', and 'e' denote the rates of contagion, incubation, infection, and recovery. 'N' represents the population of the geographical area, 'd': rate of hospitalization/ quarantine after being symptomatic, φ_1: the fraction of the symptomatic infections, and φ_2: the fraction of the quarantined/hospitalized population resulting in mortality. The details of the models could be found in (Alauddin et al., 2020; Hu et al., 2020; Li et al., 2020; Paiva, Afonso, de Oliveira, & Garcia, 2020).

$$\frac{dS}{dt} = -\frac{aS(t)I(t)}{N} \tag{8.1}$$

$$\frac{dE}{dt} = \frac{aS(t)I(t)}{N} - bE(t) \tag{8.2}$$

$$\frac{dI_1}{dt} = bE(t) - cI_1(t) \tag{8.3}$$

$$\frac{dI_2}{dt} = \varphi_1 c I_1(t) - d I_2(t) \qquad (8.4)$$

$$\frac{dQ}{dt} = d I_2(t) - e Q(t) \qquad (8.5)$$

$$\frac{dR}{dt} = (1 - \varphi_1) c I_1(t) + (1 - \varphi_2) e Q(t) \qquad (8.6)$$

$$\frac{dD}{dt} = \varphi_2 e Q(t) \qquad (8.7)$$
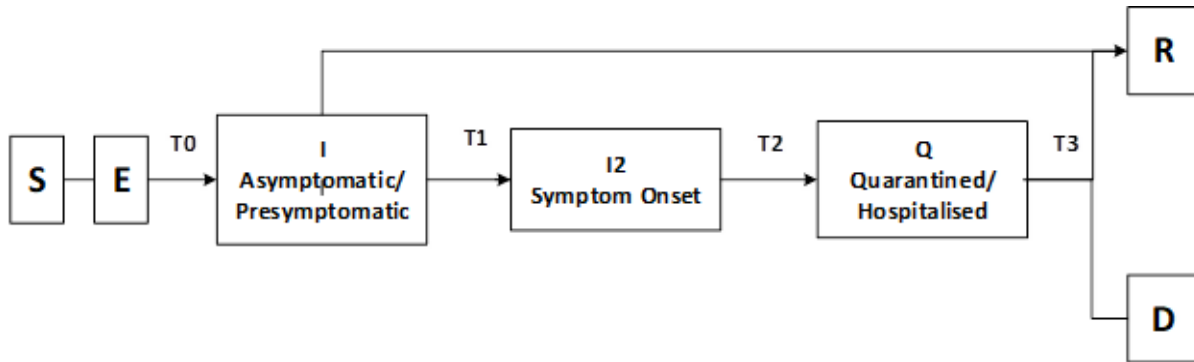
Fig. 8.1: Schematic representation of the SEIQRD model for infectious disease transmission (T_0: incubation period T_1: infection period, T_2: duration between case detection and quarantined/hospitalization, T_3: recovery period)

8.2.2: Engineering safety tools: The precautionary principle

The precautionary principle or *the precautionary approach* defines a key procedure in risk management, especially where uncertainties are difficult to quantify. It is a principle for making practical decisions under scientific uncertainty (Gollier & Treich, 2003). A precautionary decision-making approach emphasizes implementing prompt and effective preventative action even in the absence of full scientific evidence of cause and effect. UNESCO's World Commission on the Ethics of Scientific Knowledge and Technology defines precautionary principles as "When human activities may lead to morally unacceptable harm that is scientifically plausible but uncertain, actions shall be taken to avoid or diminish that harm (COMEST, 2005)".

Sandin (1999) reviewed various definitions of the precautionary principle along four key dimensions: threat, uncertainty, action, and command, as presented in Fig. 8.2. Threat refers to the nature of the imminent harm: its seriousness and irreversibility. *The precautionary principle* is about "Go slow and ask smart questions." A wide range of alternative actions, including inaction, should be examined for the severity of the potential harm along with the consideration of the moral implications.

1. **Threat/ Harm:**
   - Can be evaluated by creating and evaluating credible scenarios.
   - The credible scenario can be analyzed using risk assessment techniques

   What to consider for the threat:
   A. Magnitude:
      Is it serious?
      What's the cumulative impact?
      Would risk assessment help with the analysis of the scenario?
   B. What Kind:
      Is it irreversible?
      Is it easily avoidable?
2. **Uncertainty:**
3. **Action**

   Action taken ∝ *Seriousness of the potential harm*

   Questions to ask for action/inaction:
      Just wait, or can we use science to mitigate it?
      Can we use science to anticipate?
      What risk reduction strategies should be employed?
      Can we use science to make our system resilient towards the threat?
      Can we do something to reverse it?
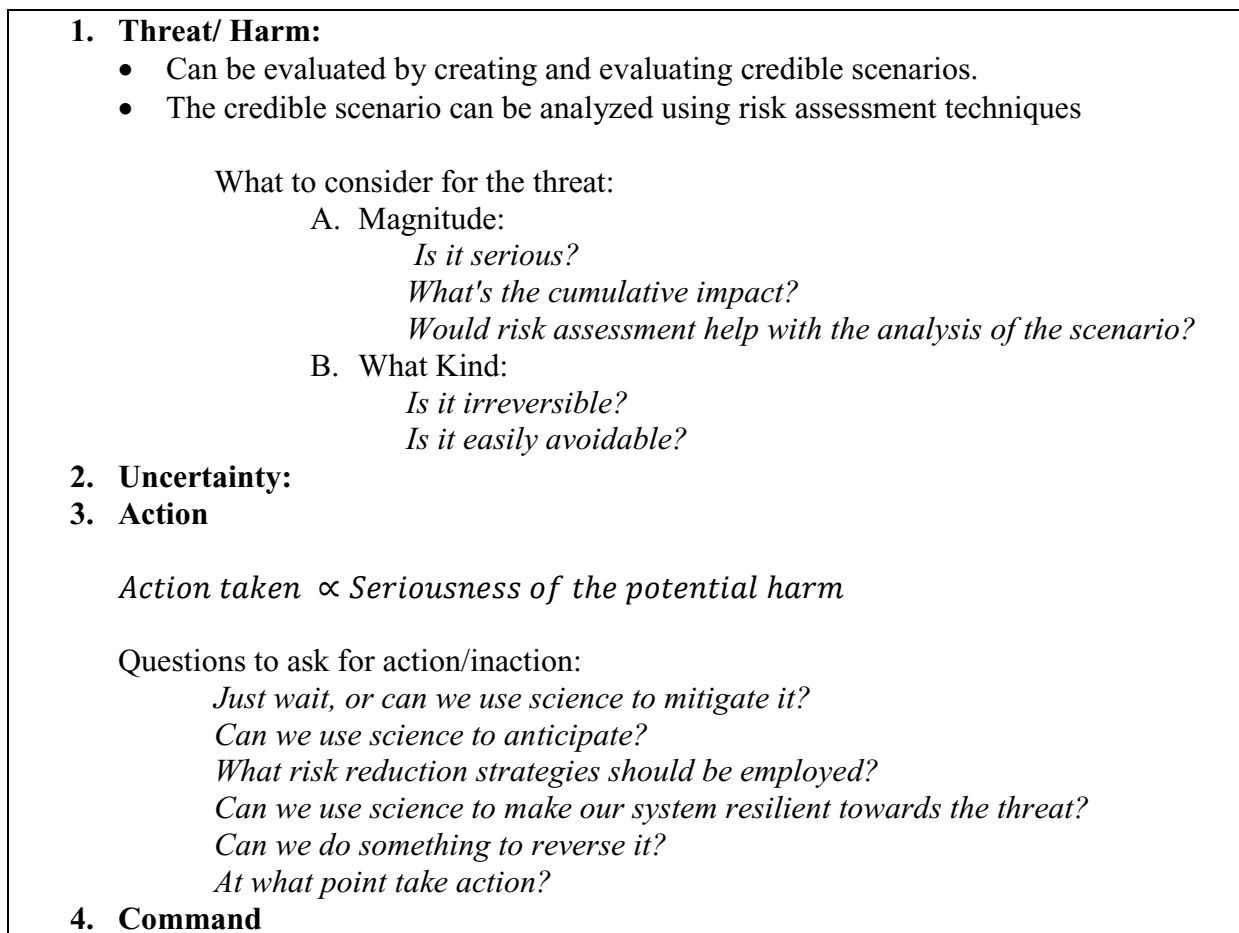   *At what point take action?*
4. **Command**

Fig. 8.2: Dimensions of precautionary principles (Sandin, 1999)

8.2.3: Engineering safety tools: The *ALARP principle*

The *ALARP (as low as reasonably practicable*) approach is based on risk-informed and cautionary thinking. The ALARP principle states that risk-reducing measures should be implemented, provided that the costs are not grossly disproportionate to the benefits earned (Pike, Khan, & Amyotte, 2020). This usually applies to the tolerability region, which is the region between intolerable and accepted risk levels. The risk should be reduced, or the activity must be discontinued if it exceeds the maximum tolerable level (Pike, Khan, & Amyotte, 2020). All critical words in ALARP: 'low', 'reasonably', and 'practicable' are relative terms with no standardized values. Risk acceptance is a complex process influenced by several factors such as the order of risk, the extent of societal participation, and corresponding regulations and guidelines.

Fig. 8.3 outlines the pandemic risk assessment using Engineering Safety tools such as the PP and the ALARP. The precautionary approach has been examined to estimate the pandemic's risk with and without implementing risk-reducing measures. The enforcement of distinct risk-reducing measures, including no measure, has been evaluated using the ALARP approach. We have employed the SEIQRD model for the quantitative analysis, i.e., calculating the number of newly infected cases, hospitalization, recovered, and mortality due to the pandemic. The randomness in the model parameters, e.g., incubation, infection, and recovery periods, has been captured using the Monte Carlo simulation. Finally, we have estimated the reliability of the existing healthcare facility under distinct strategies.
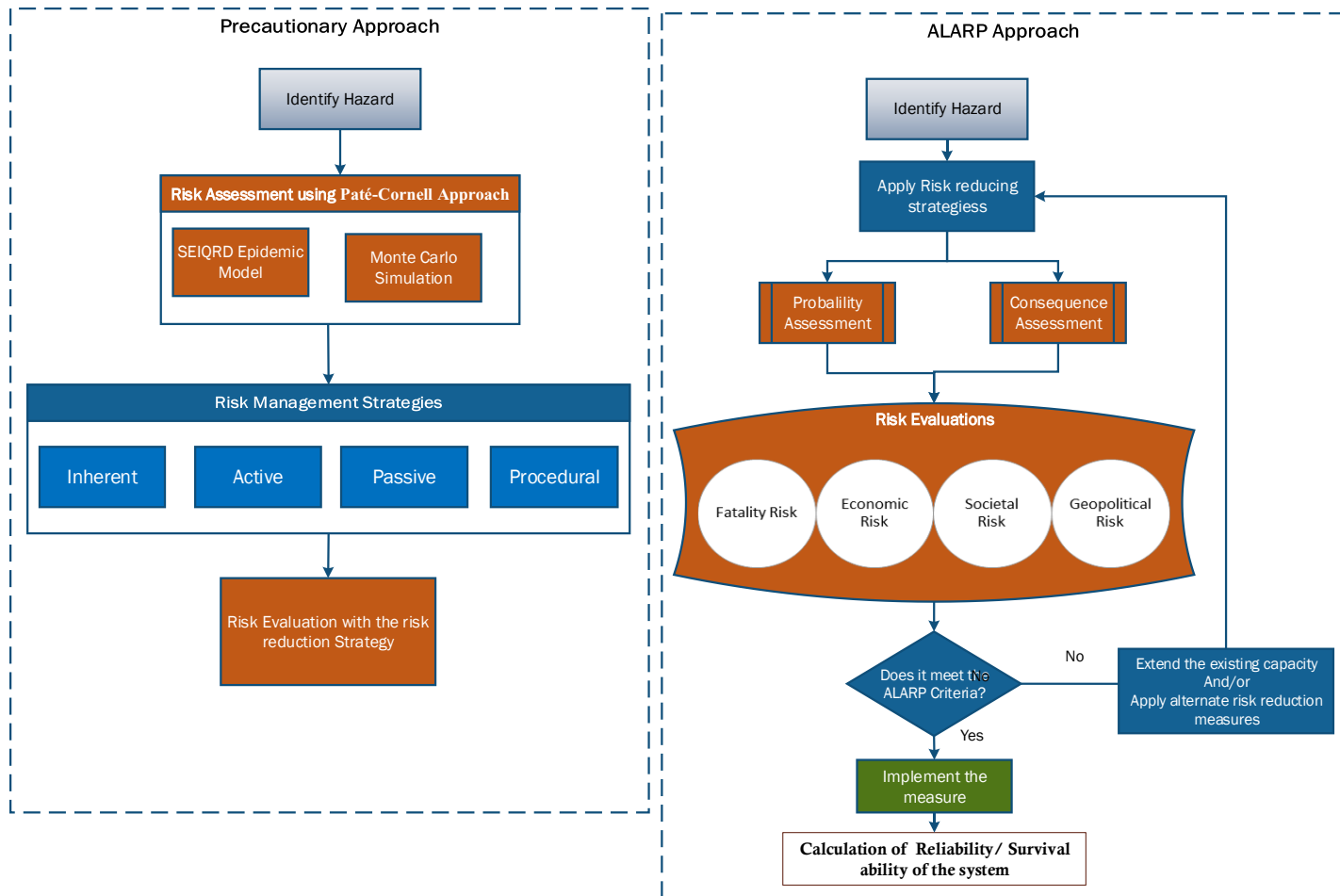
Fig. 8.3: Engineering safety based mechanistic models for pandemic risk management

8.3 COVID-19 modeling using Engineering Safety Principles

Engineering safety models have been used to study the risk management of COVID-19, a global pandemic, and severe disruption of the 21st century. The disease can lead to a range of outcomes, including no symptoms, mild illness, mental disorder, shortness of breath, sore throat, headache, myalgia, fatigue, loss of taste, fever, muscles or body aches, congestion, nausea, diarrhea, and death (CDC, 2020). The case fatality rate (CFR) of COVID-19 varies by location, the intensity of transmission, the demography, accessibility of sophisticated healthcare, and the patient's history of chronic disease. Personal hygiene (e.g., wearing a mask at public places, frequently washing

hands), social distancing, and government interventions are critical in restraining the epidemic spread of COVID-19.

The transmissibility of an epidemic is characterized by the basic reproduction number (R_0), which is defined as the average number of secondary cases generated by a primary case in an entirely susceptible population (Ferguson et al., 2005). The epidemic spreads for $R_0 >1$ and dies out if $R_0 <1$. The basic reproduction number for the COVID-19 reported by the multiple sources varies from 1.5 to 5.0. We have used R0= 2.9, the median value reported in (Liu et. al, 2020, and Alauddin et al, 2020). The average values of the incubation, infection, and recovery periods have been assigned to 5.5, 5.1, and 11.5 days, respectively.

We have studied the risk management of COVID-19 for Ontario, the most populous province of Canada, with 14.7 million people representing 38.8% of the country's population (Ministry of Finance, Government of Ontario, 2019).

8.3.1: Risk assessment of COVID-19 using PP and the SEIQRD model

The precautionary principle is fundamental in suppressing a pandemic. Fig. 8.4 presents the outline of the precautionary principles for managing the present pandemic. A pandemic outbreak contains many sources of uncertainties: strains of the virus, modes of propagation (airborne or contact transmission), the intensity of propagation (uncertainty in the R-value), rate of incubation, infection, and recovery, number of total infections, and the existence or non-existence of asymptomatic spreading. According to the precautionary principle, firm decisions need to be made to protect health in such uncertainties. The geographical region lockdown until the evidence of diminishing the disease's spread is the ultimate precautionary measure for reducing the pandemic risk. However,   it incurs severe socio-economic consequences. We have estimated the pandemic risk using the SEIQRD model under the precautionary approach.
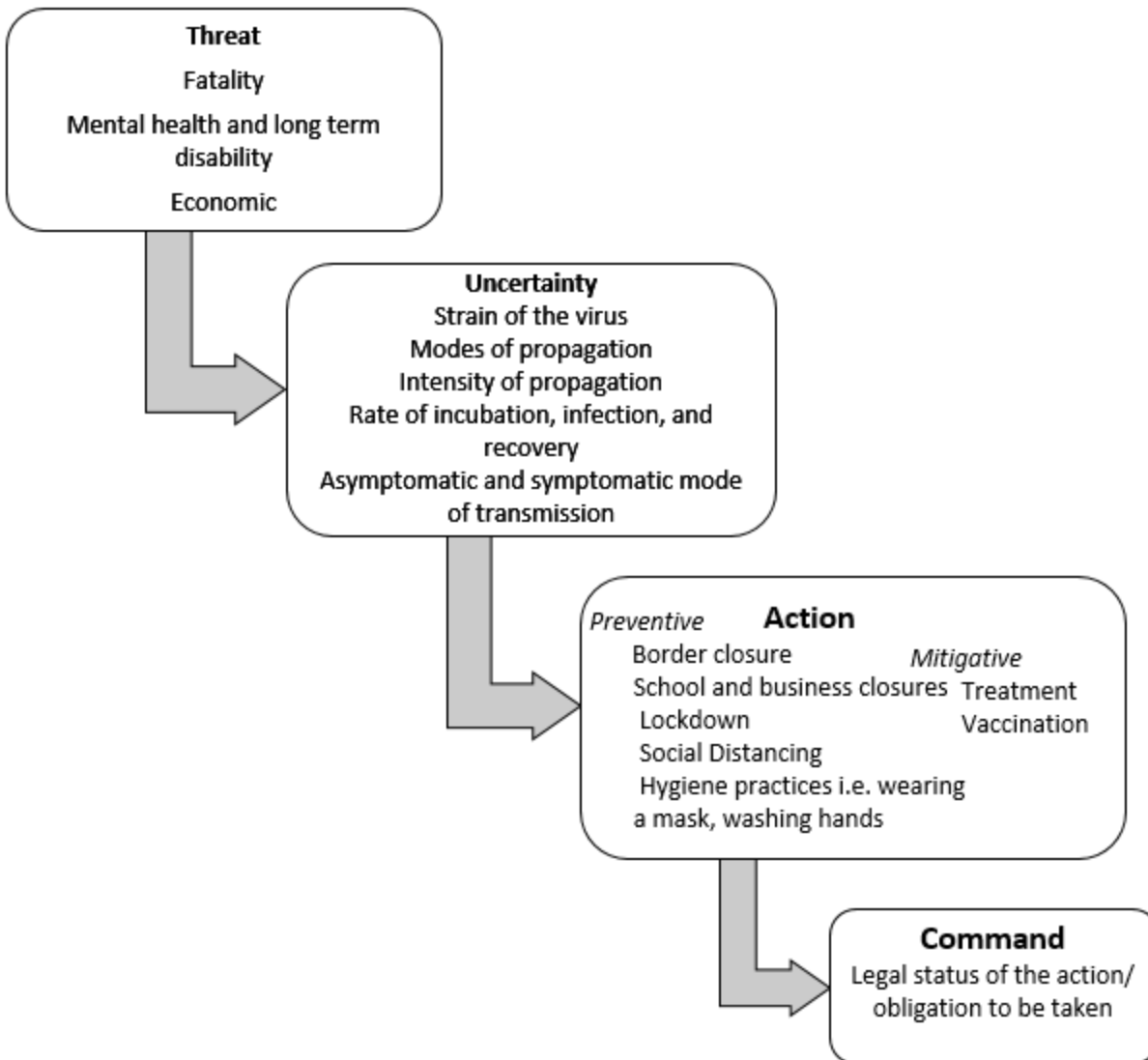
Fig. 8.4: Schematic representation of the precautionary principles for managing pandemic risk

8.3.1.1 Quantitative risk assessment of the COVID-19 pandemic

Risk assessment can be expressed in terms of answering three questions; what can go wrong, how likely is it, and what are the consequences? Paté-Cornell (1996) proposed six modes of treatment of both aleatory and epistemic types of uncertainty for risk analyses. Mode 0 is about hazard identification or multiple ways of failure of the system. Mode 1, the worst-case approach, is based on the accumulation of worst-case assumptions and provides estimated maximum loss. No notion

of probability is taken into account in this mode. Mode 2, the quasi-worst case scenario, deals with evaluating the worst possible conditions that can be reasonably expected. Mode 3 is based on a central value, e.g., the mean, the median, or the mode of the outcome. Mode 4 is based on the probabilistic risk analysis (PRA) approaches. Mode 5 displays uncertainties about fundamental hypotheses by a family of curves. Bayesian and Monte Carlo estimations are included in the tools for probabilistic risk calculation at modes 4 and 5. We have employed the SEIQRD model for the quantitative risk estimation of the COVID-19 pandemic. The Monte Carlo simulation has been used to treat randomness of the model parameters, e.g., incubation, infection, and recovery periods.

The number of daily infected cases of COVID-19 corresponding to Mode 1, 2, and 3 have been presented in Fig. 8.5. The expected and the 95 percentile values of the newly infected cases are 3.1×10^4 and 4.8×10^4, respectively. This number of daily cases could reach 7.5×10^4 in the worst-case scenario. Fig. 8.6 presents the risk considering the probability, and pandemic impact corresponds to Mode 5 of the Paté-Cornell (1996). Here, the dotted line presents the impact of the pandemic in terms of the expected values of the newly infected cases, while the solid line denotes the risk defined by the product of the impact of the pandemic and the probability of the occurrence of the impact. The probability of the occurrence of the infection has been computed using the Monte Carlo simulation of the distribution of the infections considering the randomness of the model parameters (i.e., the incubation period, infection period, and recovery period). We can also observe that the nature of the ensuing distribution depends on the relative maturity of the outbreak. Fig. 8.7 illustrates the uncertainty in the analysis. The shaded region denotes the area between 95 and 5 percentile values of the newly infected cases of COVID-19.

The aforementioned analyses are based on the SEIQRD pandemic model of the risk calculations. The analyses assume no measures were taken to restrain the spread. However, the risk is

reasonably minimized by implementing distinct risk reduction strategies, as discussed in the next section.
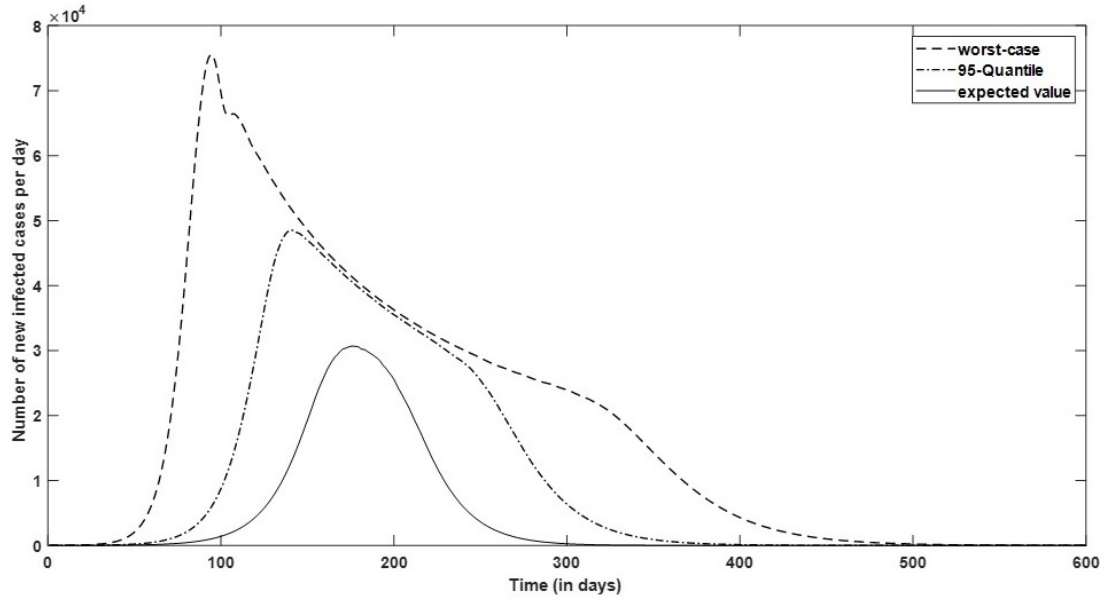
Fig. 8.5: Infection cases (Mode 1, 2, and 3) due to COVID-19 pandemic if no measures have been taken
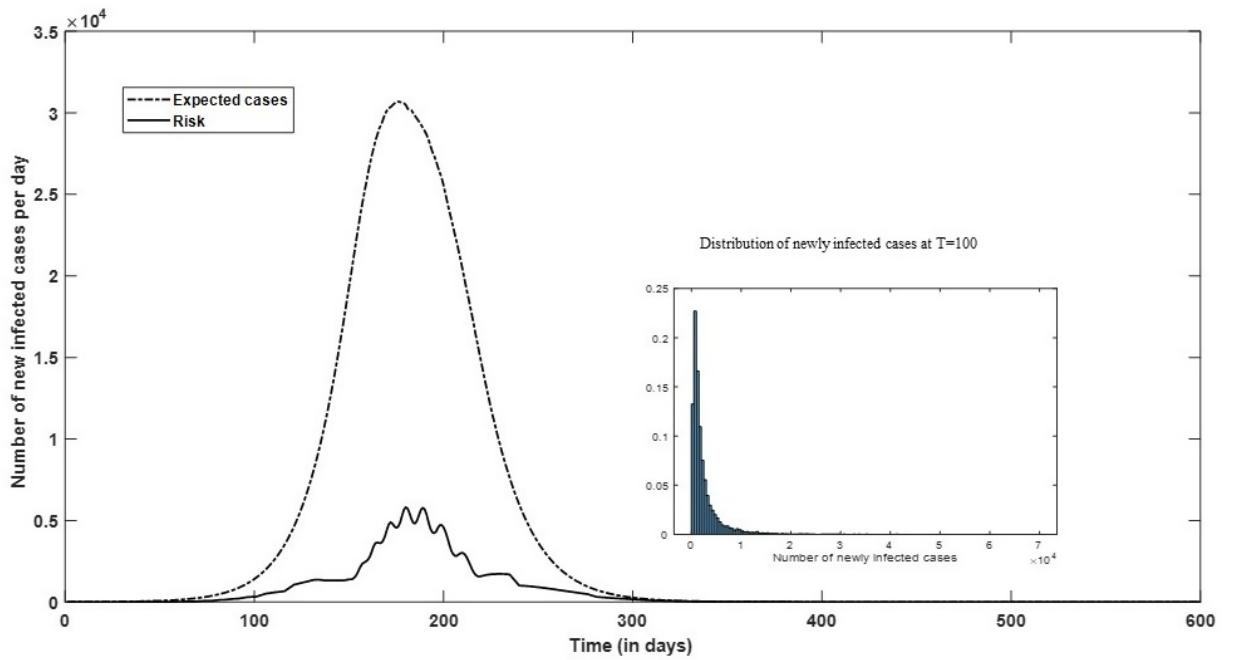
Fig. 8.6: Risk of infection (Mode 4) due to COVID-19 pandemic if no measures are taken
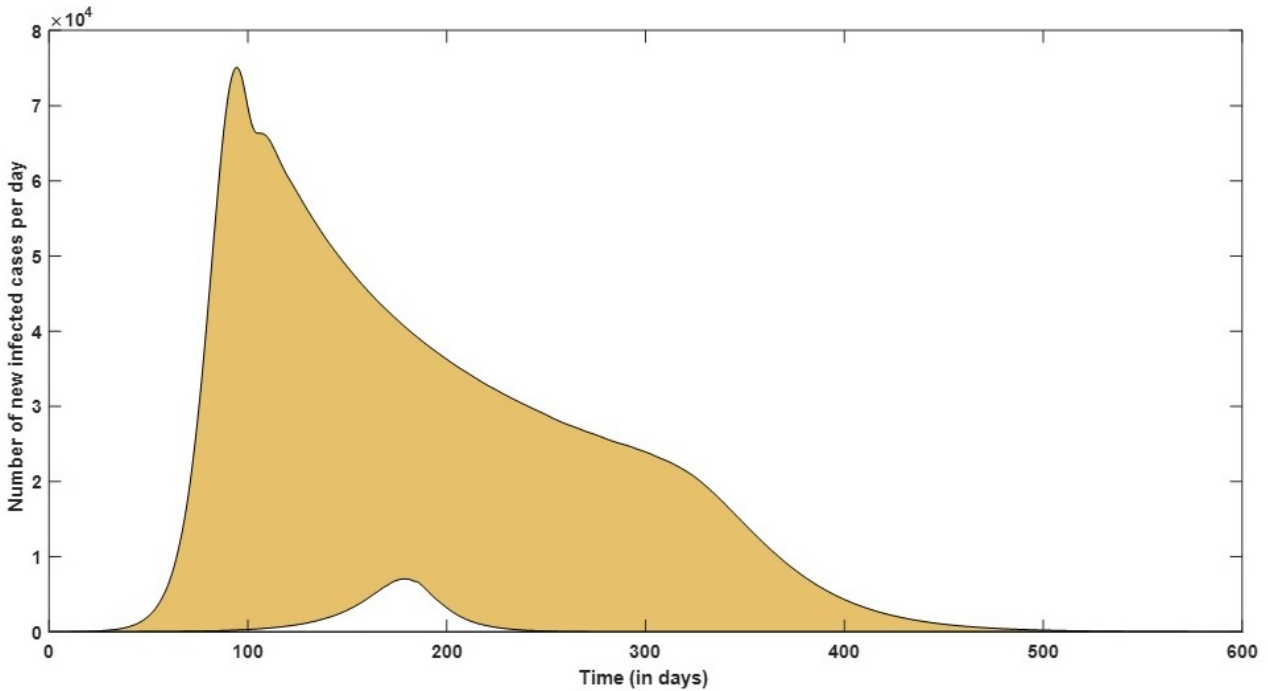
Fig. 8.7: Uncertainty in the fatality risk (Mode 5) due to COVID-19 pandemic if no measures are taken

8.3.1.2: Risk-reducing strategies of COVID-19

Following the engineering risk reduction classifications (Crowl & Louvar, 2011), the risk reduction activities for COVID-19 could be classified into four categories: inherent, active, passive, and procedural, as shown in Table 8.2. It also categorizes distinct risk-reducing measures in pre-pandemic and during a pandemic. Inherent strategies identify and implement ways to eliminate or significantly reduce the hazard. They are described by four actions: minimization; substitution; moderation; and simplification. Although inherent strategies perform well when considered early in the life cycle of industrial activity, they can be applied at any stage to reduce the risk of existing activities (Amyotte, Irvine, & Khan, 2018).

Birds and animals act as a source, reservoir, and carrier for most infectious diseases. A study reveals that 62% of all human pathogens are classified as zoonoses (Vorou, Papavassiliou, &

Tsiodras, 2007). Bats are notorious primary sources of pandemic-causing viruses such as MERS (Middle East Respiratory Syndrome), SARS (Severe Acute Respiratory Syndrome), and Nipah. A pandemic can be averted by avoiding interaction and handling of birds and animals. However, it is impractical as many people interact with them for food, fibre, livelihoods, transport, sport, companionship, and education. They can also be infectious via other transmission media, e.g., air, water, and soil, even if we avert direct contact. Another inherent strategy for preventing a pandemic is by avoiding human-to-human interactions. This is possible by changing the operational formats such as activating home delivery services, working from home, and switching to teleconferencing and virtual modes of operation. Nonetheless, the absolute interaction-free environment is highly unlikely to date due to two obvious reasons: (i) the virtual modes is not feasible for all activities and workplaces due to their reciprocative nature e.g., healthcare workers (ii) the requirement of a fraction of the workforce for the maintenance of the virtual environment. Many experts believe that an infectious disease outbreak could be wiped out if the world stands still for around the virus's survival time.

Lockdown, school and business closures, restricting large gatherings, following social distancing, putting on PPE, and hygiene practices such as frequent hand-washing are other common inherently safer approaches to pandemic risk management (Brown, Amyotte & Vanberkel, 2021). Lockdown, school and business closures, and other government regulations have other associated risks such as mental health disorders and severe economic impairments (Singh et al., 2020). These advisories entail making informed decisions on when to activate and relax various enforcements.

Contact tracing, increasing testing capacity, and quarantine of the exposed cases can be classified as administrative strategies of pandemic risk management. They are compelling in limiting the disease outbreak (Institute of Medicine, 2007). However, they must be triggered at right time to

achieve the desired outcome (Alauddin et al., 2020). A delay in detecting infected cases leads to a delay in the mitigative actions that escalate the risk. Hygiene practices such as frequent washing of hands and refrain from face-touching are other proven active measures for suppressing the disease if exposed to the Coronavirus.

Unlike active strategies that require event detection and device actuation for their functioning, passive engineering safety strategies comprise barriers that do not need activation to accomplish their intended functions. Bolstering immunity either by changing lifestyle or achieved through herd protection is an effective passive strategy for reducing the pandemic risk. The shield at cash and other counters is another example of a passive control mechanism of restraining the disease spread. The passive strategies require long-term planning. The present outbreak can help upgrade our passive control systems for reducing the risk of future infectious diseases.

Providing sophisticated treatment to infected people is a procedural method for mitigating a pandemic risk. The existing healthcare facilities might need to be extended to meet the demands of treating a large number of infected cases. Thoughtful decisions have to be made to mobilize resources and aid preferential treatment to vulnerable groups in case of limited availability. The other effective procedural strategies include awareness about the situation, special attention and guidelines for the vulnerable groups, e.g. elderly and chronic patients, peer pressure, and police intervention for following the procedure.

Categorizing strategies into inherent, active, passive, and procedural is subjected to the focus of the study. Social distancing, hygiene practices, and other enforced regulations such as lockdown, school and business closures, and restricting large gatherings are inherent risk reduction measures for a susceptible person. However, these factors can be documented as procedural measures for alleviating the pandemic risk to a community if the virus is already present in the community.

Table 8.2: Categorization of risk-reducing strategies for COVID-19 pandemic

Type of measures/ barriers	Stage and risk reduction strategies	Type of risk reduction strategies	Nature and implementation of the risk reduction strategy	Remarks
Preventive	**Pre-pandemic** • Avoid direct contact/ interaction/ handling animals	*Inherent*	-	Extremely difficult to implement. Many known mammals play a vital role in human life but act as potential virus sources and/or carriers. For instance, MARBURG 1967 (bat) EBOLA 1976 (bat) SARS 2002 (bat) SARS 2012- (bat) SARS-CoV-2 2019 (presuming bat) MERS 2010 - (Camels) H5N1 (Bird flue) 2003- (Chicken) H7N9 (Bird flue) 2013- (Chicken)
	During-pandemic: • Avoid physical interaction with others • Activate work from home strategy and home delivery services	*Inherent*	Administrative recommendation that requires to be practiced by individuals and organizations	Effective mechanisms to prevent a pandemic. However, all individuals and operations cannot go online. Besides, there is a possibility of defaulters depending upon the level of administrative action (recommendation, requirement, and its enforcement)
	During Pandemic: • Enforcing lockdown • School and business closures • Restricting large gatherings • Frequent hand washing/sanitizing/ refrain from face touching	*Active*	Administrative recommendation that requires to be practiced at individual and community level	Effective mechanism in minimizing the pandemic impacts. However, it is challenging to enforce and monitor enforcement. They incur significant economic loss

	• Social distancing Avoiding crowded places/ public transport	*Inherent*		
	During Pandemic: • Vaccination	*Inherent*	Administrative recommendation that requires to be followed by individuals	The most effective strategy. It provides the fastest way to minimize the pandemic impact provided Vaccine is available and accessible to all.
	During Pandemic: • Redesign/installation of safety layers at the interactive systems, e.g., shield at cash and other counters	*Passive*	Engineering	An effective strategy to minimize the disease spread. However, it requires proper planning and execution.
	During Pandemic: • Self-isolation • Wearing a mask/ PPE • Good Hygiene practices • Surface Cleaning	*Procedural*	Administrative recommendation that requires to be practiced at individual and community level	The efficiency of the strategy is dependent on individuals to follow the best practices.
	During Pandemic: • Contact Tracing • Rapid Testing • Awareness about the situation and safe handling procedures • Peer pressure and police intervention for following procedures • Special attention and guidelines for the vulnerable groups	*Procedural*	Administrative	It requires significant resources to enforce the measures.
	During Pandemic: • Immunity	*Passive*	Achieved through herd protection, genetics or use of diets to strengthen the immune system.	This is an effective passive strategy; however, it is highly variant depending upon the individual's immune system.
Mitigative	**During Pandemic:**	*Procedural*	Administrative	

				It requires decisions to activate the strategies effectively and mobilize the resources.
	• Quarantine of exposed cases • Treatment • Extending healthcare systems/ hospitals/ workers/antidotes			Requires long-term planning The prevalent outbreak can be used to upgrade the healthcare systems to respond well in future outbreaks.

8.3.1.3: Risk management of COVID-19 using risk-reducing measures

Distinct government regulations and individual responses can minimize the risk of a pandemic. Limiting gathering sizes, closure of nonessential businesses and schools, and emergency lockdown have a decisive role in controlling the epidemic spread. Lockdown is the most effective measure for reducing risk. However, prolonged strict lockdown can cause compromised mental health and severe economic impairment. We have modeled the lockdown as a precautionary approach. Fig. 8.8 demonstrates the effect of lockdown after one week of the first mortality reported. The lockdown will reduce the peak to less than 10 newly infected cases as opposed to 3.1×10^4 if no measures are taken. The variation of the value due to randomness is presented using the shaded area in Fig. 8.8. The timing of the enforcement and relaxing of the lockdown is crucial in restraining the epidemic risk (Alauddin et al., 2020).
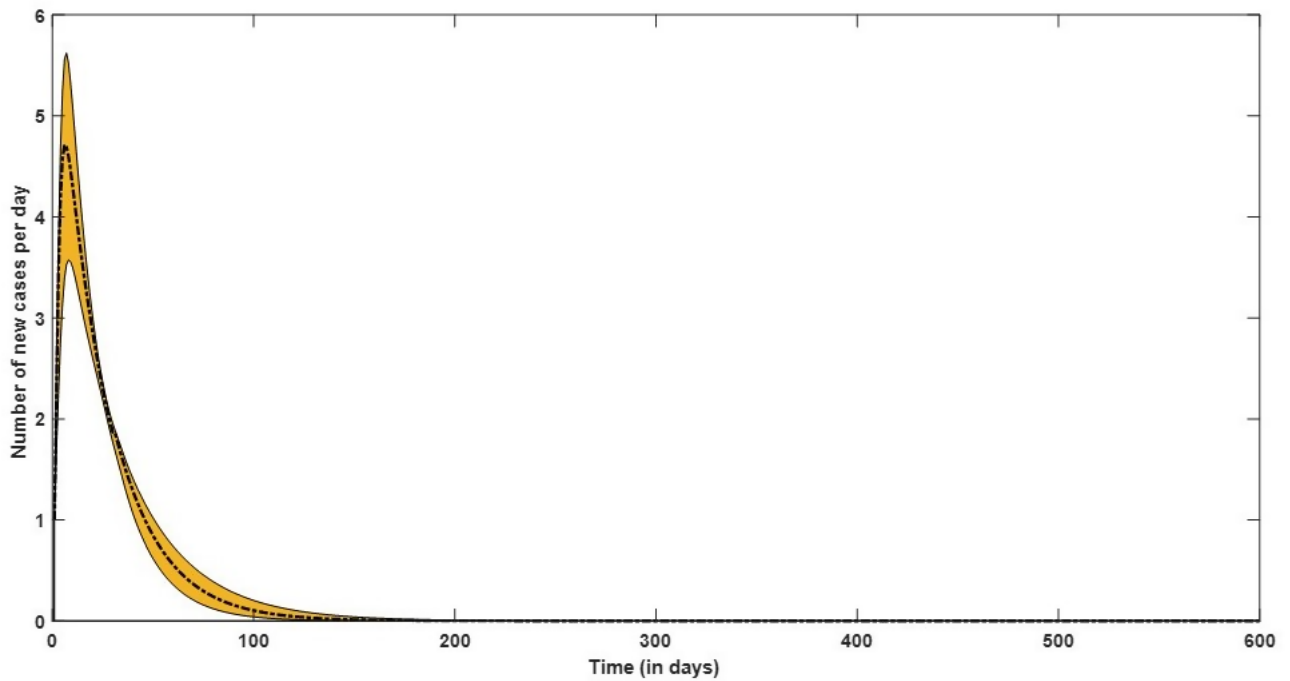
Fig. 8.8: Newly infected cases (per day) of COVID-19 pandemic under the lockdown

The results present the effectiveness of the lockdown (and other interventions in the forthcoming sections) in reducing infections and fatality in relative terms. The estimates were not corrected for other potential confounders' effects, for example, wearing a mask, hygiene practices, and following voluntary social distancing that could have contributed to reducing the disease spread in addition to the observed interventions. Our study also does not explicitly consider the other key factors such as the scale of testing, contact tracing, and imperfections in case isolation, demographics, heterogeneities in contact patterns, and spatial effects. Besides, the results were not adjusted for fatalities arising from the interruption in health services for chronic disease. Many surveys highlighted the partial or complete disruption of healthcare for hypertension, diabetes-related complications, cancer treatment, and cardiovascular emergencies due to the newly imposed regulations during the COVID-19 pandemic.

The effectiveness of these measures depends on the context of implementation, such as the presence of other NPIs, country demographics, the economic status, degree of compliance of the population and societal impact, comorbidities, overall risky environment, and country vulnerability to biological threats. The same NPI can result in different outcomes in different countries, even different parts of a country. Application of expertise from similar outbreaks in the past could be conducive to the credible estimate of the trajectory and slowing down the spread (Goudarzi, 2020, March 23). The expertise from the past outbreaks e.g., the 2003 severe acute respiratory syndrome (SARS) outbreak in Singapore, and experience with the 2015 Middle East respiratory syndrome (MERS) outbreak of South Korea led to an immediate fruitful response to the COVID-19. The economic support by emergency funding to the poor and needy populations for alleviating the financial burden is also helpful in effective closures and lockdown.

8.3.2 Event tree analysis of the COVID-19 pandemic

A pandemic can cause socio-economic damage, compromised mental health, and mass mortality. Many preventive and repressive or mitigating measures have been explored to minimize the negative consequences of infectious diseases. The term 'prevention' refers to measures taken to prevent the occurrence of an unwanted event while 'repression' translates to the measures taken to mitigate the consequences of the undesired event. Repressive barriers are put in place to avert, mitigate and minimize the adverse effects of the central event (Lindhout & Reniers, 2020).

Fig. 8.9 depicts the impact of the epidemic on an infected individual as well as on the community. The end states have been divided in two subgroups: risk to an infected person (including safe and death as the scenario), and risk to the community with many scenarios such as safe, low risk, moderate risk, high risk, very high risk, and exceedingly high risk. Lockdown, school and business closures, self-isolation, and social distancing, significantly reduce the risk. Fig. 8.9 also illustrates

that the asymptomatic spread can be catastrophic if not mitigated properly. A detailed analysis of the event tree and bow-tie analyses of COVID-19 for subsystems can be found in (Brown, Amyotte, &Vanberkel, 2021).

 Super-spreading incidents and multiple infections from a single infected individual were the key drivers of the COVID-19 transmission (Frieden & Lee, 2020). Some of those events include the Biogen meeting (Weintraub, 2020), the Caul's Funeral Home at St John's (Courage, 2020), the White House Event (BBC, 2020), the Cluster of Bars in Hong Kong (Danmeng & Jia, 2020), and the Church Choir Practice in Washington (Williams, 2020). Asymptomatic and pre-symptomatic infections are the greatest means of spreading diseases. According to a reprint (Goyal, Reeves, Cardozo-Ojeda, Schiffer, & Mayer, 2020), about 62% of super-spreading COVID-19 occurred through pre-symptomatic transmission. Interventions such as lockdown, school and business closures, and limiting gatherings are central in preventing super-spreading events.

Lindhout and Reniers (2020) presents a risk management framework considering the root cause of the long-term effects of a pandemic outbreak. Many studies proposed temporal and spatial segregation measures to restrain a pandemic. Government interventions such as school and workplace closures, stay-at-home orders, and a ban on large gatherings will cause a community-wide contact rate reduction. However, these interventions cannot be imposed for a longer duration due to high incurred costs and other associated risks. Individual practices and societal responses are central to the effectiveness of these measures. Self-imposed measures such as wearing a mask at public places, voluntary social distancing, and handwashing are vital in preventing subsequent waves of an outbreak.

An event tree presents the known consequences of an abnormal event. Fig. 8.10 shows the Event Tree model of distinct risk reduction strategies of a pandemic. The risk will be negligible if

immunity is achieved either through natural, i.e., herd immunity or vaccination. However, it takes several months following the outbreak. Government interventions such as lockdown, school and business closures, restricting large gatherings, and extending healthcare systems help in restraining a pandemic. Corporates and employers can assist in controlling the risk by transforming operational formats, such as enabling home delivery services, working from home, and switching to a virtual mode for meetings. The individual responses: following social distancing, wearing a mask, and hygiene practices efficiently repress a pandemic's spread.

The efficacy of all barriers is not alike; some are more prone to failure due to their distinct nature, porosity, constraints, and degradation characteristics. For example, the individualistic-based measures, e.g. social distancing, washing hands, could be weakened due to people's complacent nature, especially if the outbreak persists for a longer duration. Likewise, the lockdown cannot be imposed for a prolonged time due to its severe economic consequences. The measures are interactive; the effectiveness of each measure depends on the others. For instance, the success of government regulations for restraining a pandemic is vastly influenced by social responses and individual practices. Multiple strategies improve the reliability of disease transmission barriers.
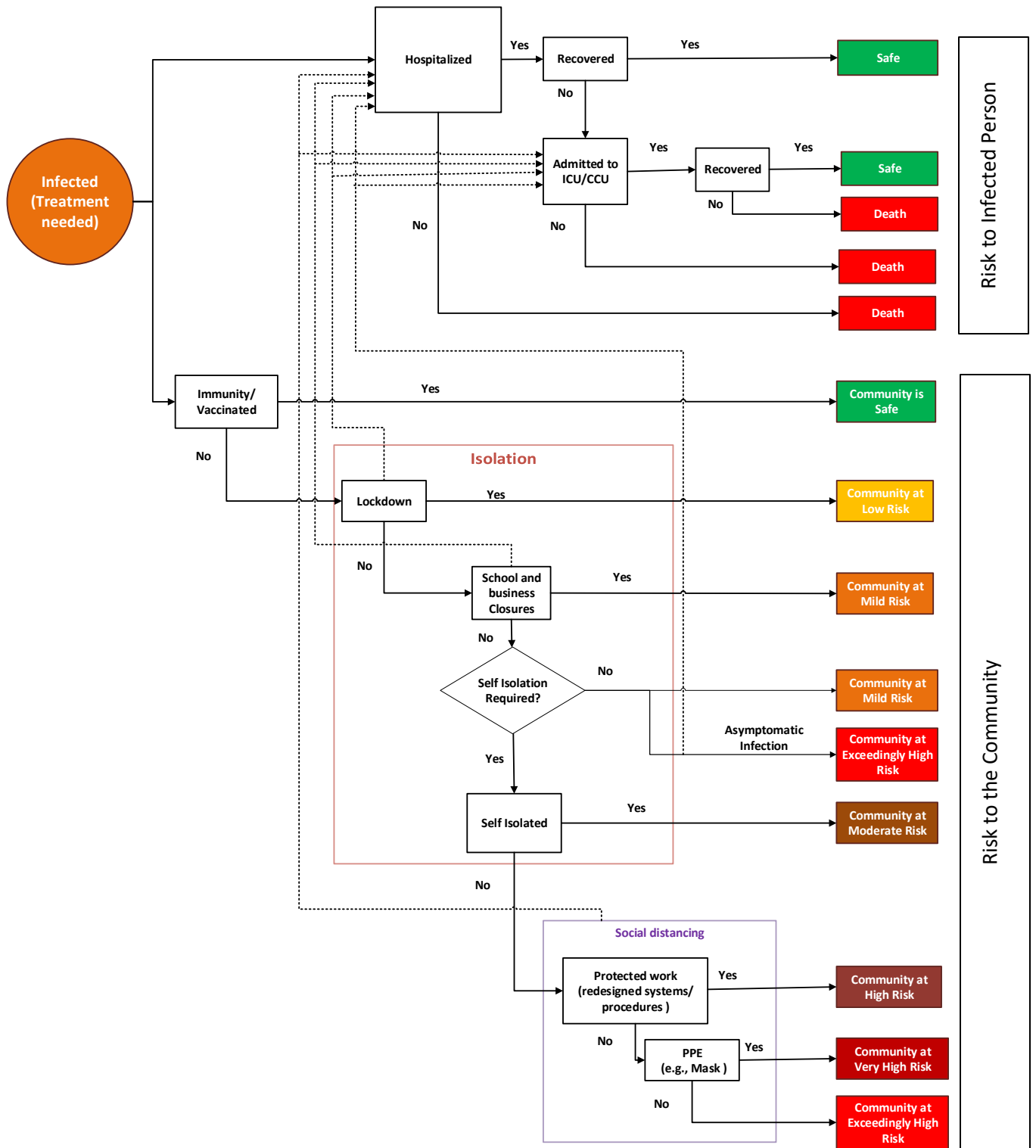
Fig. 8.9: Impact of a pandemic on an infected person and the community

Fig. 8.10: Event Tree model of distinct risk reduction strategies of a pandemic

Non-pharmaceutical interventions also play crucial roles in allocating acute and critical care beds. Fig. 8.11 shows the estimation of the availability of acute and critical care beds during a pandemic. The vulnerability of infected individuals rely on their health history and the availability of sophisticated treatment, which depends upon the following factors-

1. The capacity of the healthcare system

2. The stage at which a person is getting infected. This is because the existing beds might be occupied by other patients if the person is being infected at a relatively mature stage of the outbreak.

3. The intervention(s) enforced

The temporal variation of the hospitalization status and the new cases due to the COVID-19 pandemic under distinct regulations (i.e. no measures, school and business closures, and lockdown) has been presented in Fig. 8.12. We have assumed that 25% of the infected persons are home quarantined. We can observe that the healthcare systems would be exhausted quickly if no measures were taken (Fig. 8.12A). However, the existing healthcare system would suffice under the schools and business closures (Fig. 8.12B) and lockdown (Fig. 8.12C). Table 3 and Table 8.4, respectively, present the consequences if someone is infected at **T=200th** and **T=550th** day since the outbreak. We have assumed that Ontario's initial health care system has **10000** acute care beds for COVID patients, with **1000** beds available for critical and intensive care (Barrett et al., 2020). However, Ontario's government has been significantly expanding the healthcare capacity in preparation for the COVID-19 outbreak.

 A simplified Event Tree diagram for the potential risk when infected on the 200th day is presented in Fig. 8.13. Here, natural healing, acute care, and intensive care are the barriers against fatality due to COVID-19. The success of a barrier represents the availability of the barrier and successful

recovery resulting from the treatment. The facilities' allocation depends on the healthcare capacity, the enforced intervention, and the stage at which one has got infected, as discussed earlier. The probability of the bed allocation on the 200th day of the outbreak is scant that results in a high likelihood of fatality if no measures are taken. However, the bed allocation probability is virtually 1 (i.e., guaranteed bed allocation) with a 97% probability of safe recovery if restrictive school and business closure measures or the lockdown are enforced. The risk of fatality has been calculated assuming a 90% recovery rate of the acute care and 70% recovery of critical care systems. The admission and recovery rate of intensive care is a complex function of many factors, including age, gender, geography, treatment, and comorbidities, i.e., the overlap of multiple medical conditions. That might be a reason for the lack of consensus in reporting the proportion of ICU admissions. Abate, Ali, Mantfardo, and Basu (2020) reported the rate of ICU admissions of 32% (95% CI: 26 to 38, 37 studies and 32, 741 participants). The other commonly reported values of ICU admission include 5% (Guan et al., 2020), 16% (Grasselli, Pesenti, & Cecconi, 2020), and 20% (Baker et al., 2020) of all hospitalized patients.

An early report from China stated a mortality rate of 80% in ICU; however, this mortality rate dropped to one-third and improving over time (Abate et al., 2020; Launey et al., 2020). We have assumed a 10% admission rate to ICU and a 30% mortality rate of intensive care units in this computation. We have not quantified the recovery from natural healing due to data unavailability in this regard.

Table 8.3: Risk to the infected person if infection at the 200th day of the outbreak with an acute care bed capacity of 10000 and ICU bed capacity of 1000.

Assuming a 90% recovery rate of acute care and 70% recovery rate of critical care systems

a. *If no measures enforced:*

New cases seeking acute care = 19170
Occupied bed/ (old cases)= 116917 (exceeding bed capacity)
Probability of allocation of bed= 0
Probability of safe recovery= $0.00x.90 + 0.00x\ 0.10x0.00x0.70 = 0.00$
Probability of Death due to unavailability of the acute care=1
Probability of safe recovery $= 1 - 0.00 = 1$

b.   *School and business closures:*

New cases seeking acute care = 47
Occupied bed/ (old cases)= 255
Available beds for allocation=10000- 255= 9745
Probability of allocation of bed $= 1$
Probability of safe recovery= $1x.90 + 1.00x\ 0.10x1.00x0.70 = 0.97$
Probability of Death= $1 -$ Probability of safe recovery $= 1 - 0.97 = 0.03$

c.   *Lockdown:*

New cases (based on the most probable value) = 0
Occupied bed/ (old cases)= 0
Available beds for allocation=10000- 0= 10000
Probability of allocation of bed $= 1$

Table 8.4: Risk to the infected person when infection at T=550 with an acute care bed capacity of 10000 and ICU bed capacity of 1000

	No measures	*School and business closures*	*Lockdown*
New cases seeking acute care	0	2155 1652	0
Occupied bed/ (old cases)	0	8870 9235	0
Probability of allocation of bed	1	$\frac{10000-8870}{2155}= 0.46$	1
Probability of safe recovery	1	$0.52x.90 + 0.52x0.1x\ 0.70 = 0.45$	1
Probability of Death due to the pandemic	0	0.55	0
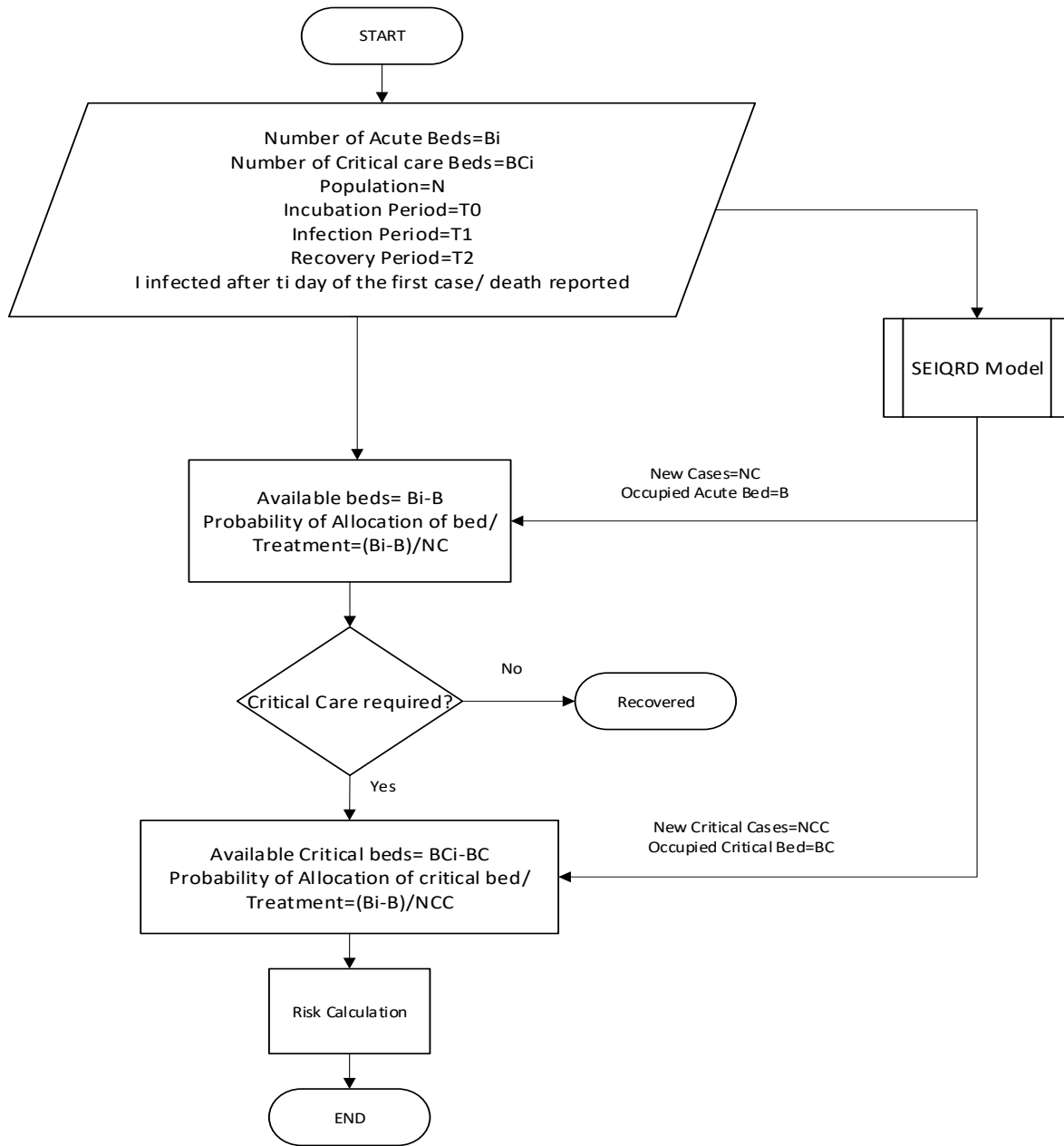
Fig. 8.11: Flowchart for estimating availability of acute and critical care beds during a pandemic
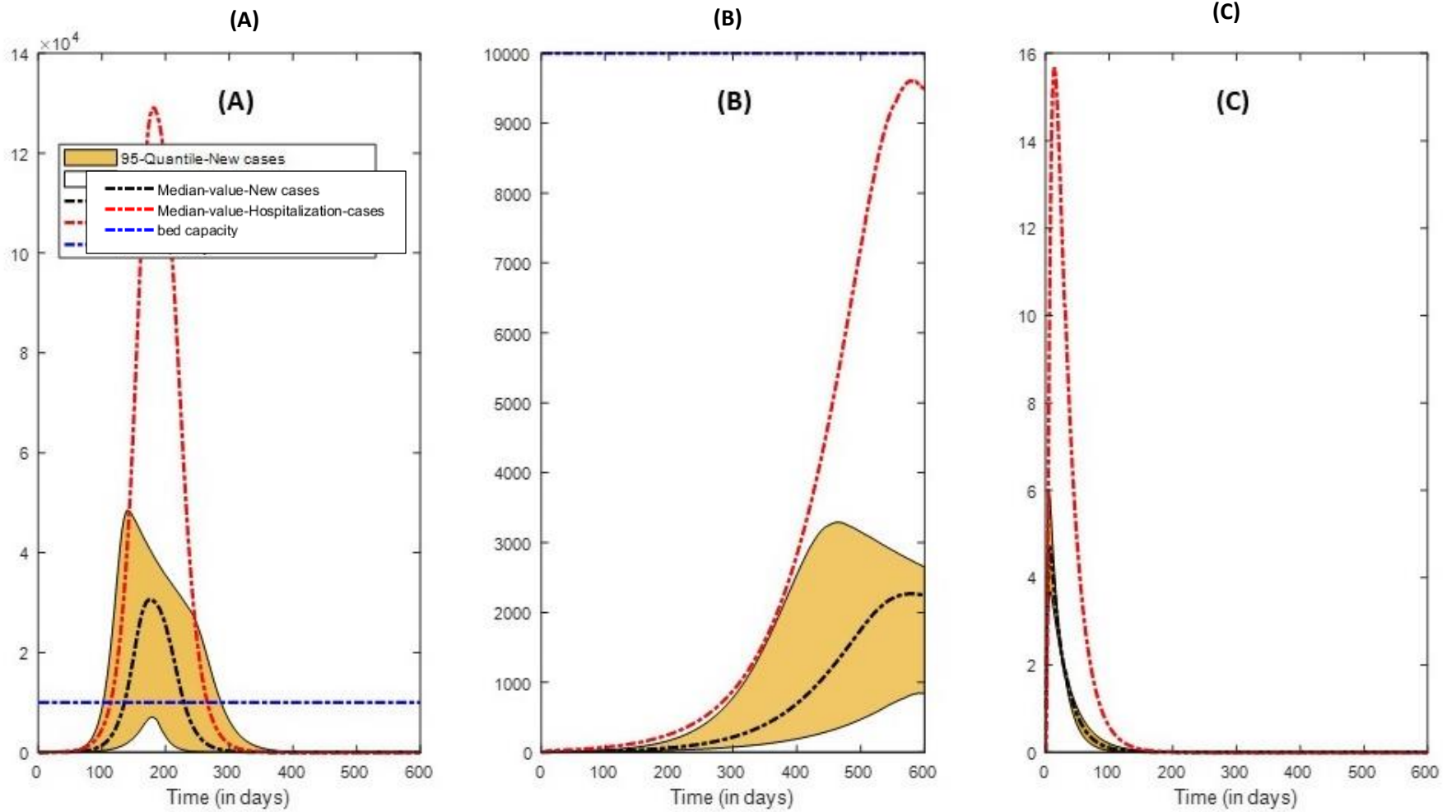
Fig. 8.12: Estimated infected cases due to the COVID-19 pandemic under distinct measures
A. No measures, B. School and business closures, C. Lockdown
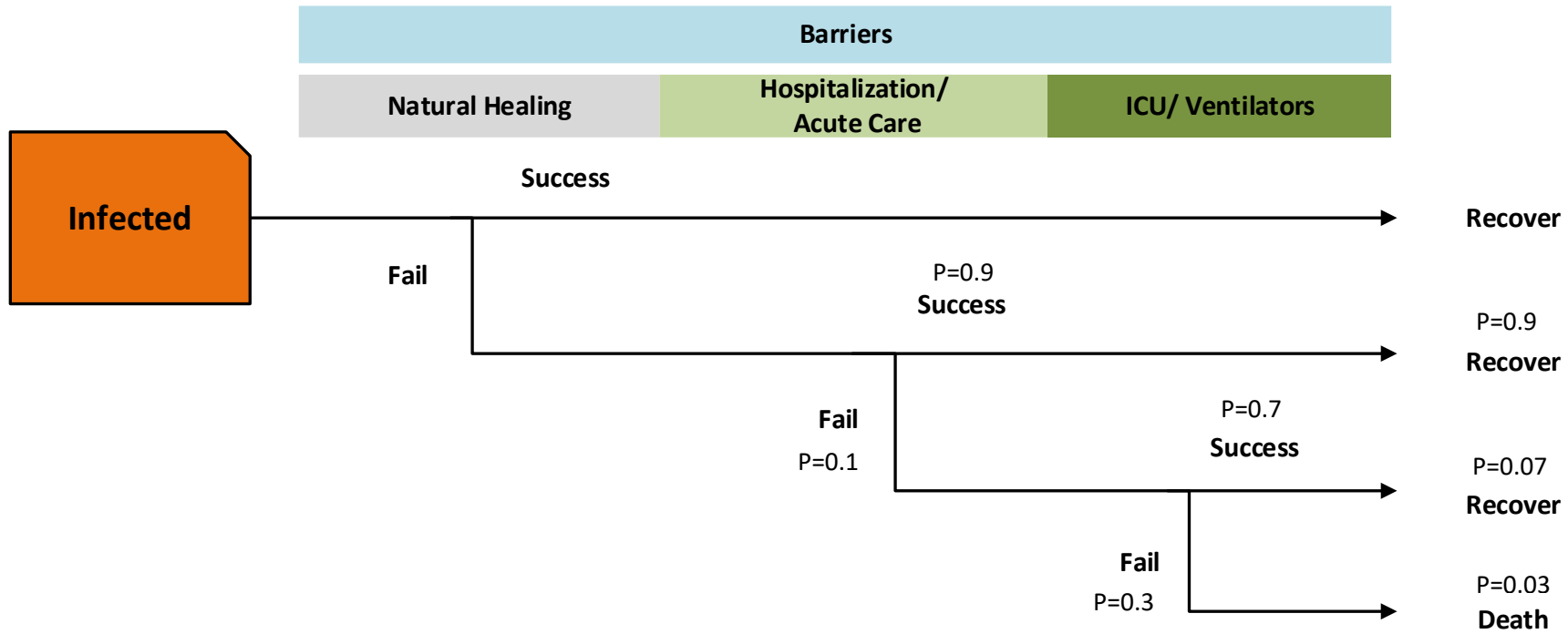
Fig. 8.13: Event tree analysis for risk to an infected person **at T=200th** day of the outbreak with schools and business closures in effect

8.3.3 Risk analysis using ALARP

The ALARP principle states that risk-reducing measures should be implemented, provided that the costs are not grossly disproportionate to the benefits earned (Pike, Khan, & Amyotte, 2020). We have assumed that the stricter the regulation, the higher will be the economic infliction.

Fig. 8.14 shows the ALARP representation of the tolerable risk of the COVID-19 pandemic for Ontario. The approach sets an upper limit above which the risk must be reduced and a lower limit below which the spent resources yield a marginal reduction in the fatality risk. The region could be divided into acceptable, tolerable, or unacceptable where the cases surpass healthcare capacity. With no intervention, the healthcare would not suffice the 12.8×10^4 cases requiring medical care on the 187th day of the pandemic. The fatality risk could be minimized either by extending the healthcare capacity or by enforcing interventions. The model predicts 9600 hospitalization cases for several weeks with school and business closures. This number drastically reduces to 15 hospitalization cases for a few days with the lockdown. Imposing interventions and expanding the healthcare capacity would be a practical approach to addressing a pandemic.
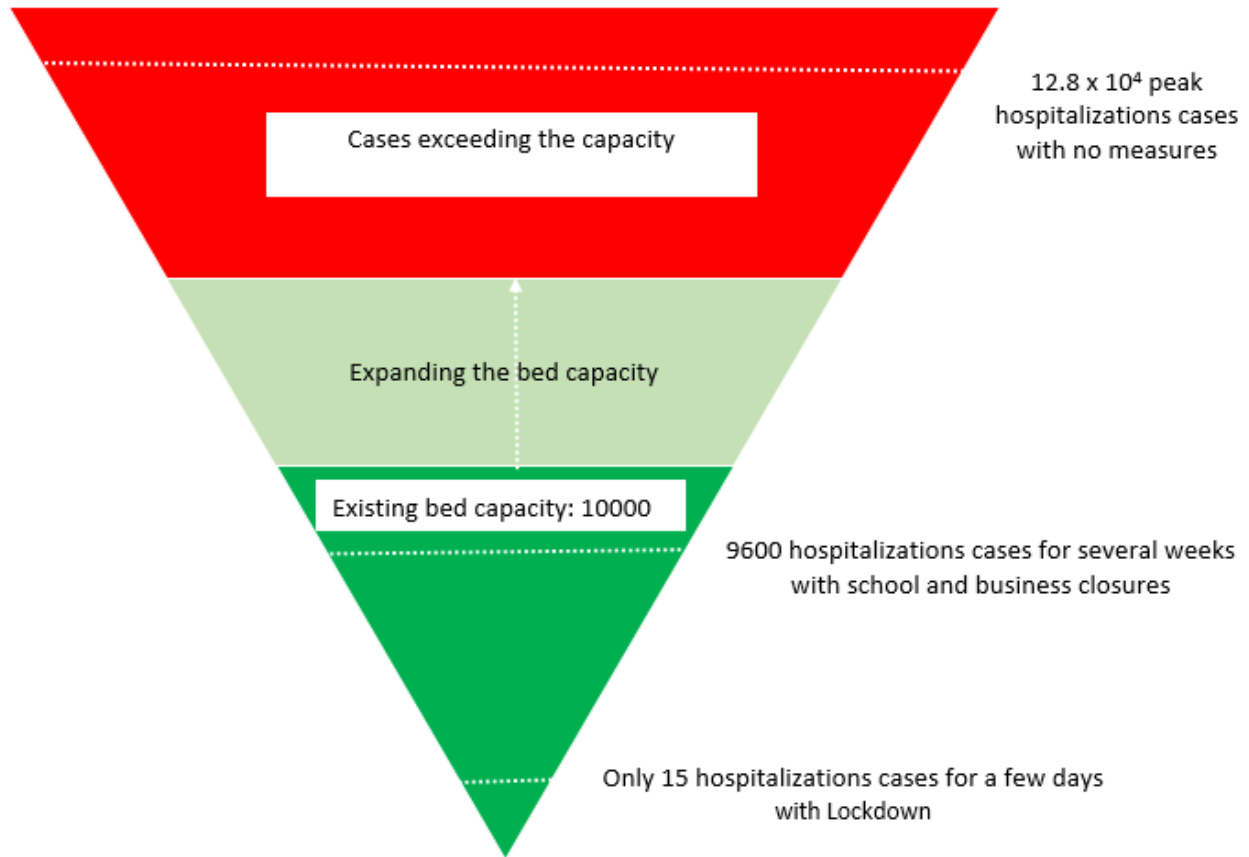
Fig. 8.14: The outcome of the ALARP based implementation for the risk management in COVID-19

8.3.4: Reliability/ Survival analysis with the existing healthcare systems

The worst possible outcome of the disease for an infected individual would be death. The infected people could be recovered if they avail of the sophisticated treatment. Healthcare accessibility depends upon the capacity of the healthcare system, the stage of the outbreak at which one is infected, and imposed intervention(s) (Section 8.3.2).

Figures 8.15-8.17 present the survival analysis for the COVID-19 pandemic with the existing healthcare facilities (critical care beds: 1000) under distinct scenarios with no measures, schools and business closures, and lockdown. The survival estimates are based on the Monte Carlo simulation to capture uncertainties in the number of infection cases due to randomness in the

incubation, infection, and recovery periods. This is represented by the area under the probability distribution of the cases requiring critical care on a given day. Similarly, it is also represented by the value of the cumulative probability distribution of the cases requiring critical care. These survival computations are based on the accessibility of critical care; the true values would be lower because of the fractional recovery rate (less than 100%) of the treatment.

The most probable estimates indicate that the treatment is not accessible to infected individuals for most of the peak durations if no measures are enforced (Fig. 8.15). The survival ability under the existing healthcare system would be negligible in this case. The corresponding survival ability values for the schools and business closures and lockdown with the existing facilities are 80% (Fig. 8.16) and 100% (Fig. 8.17).
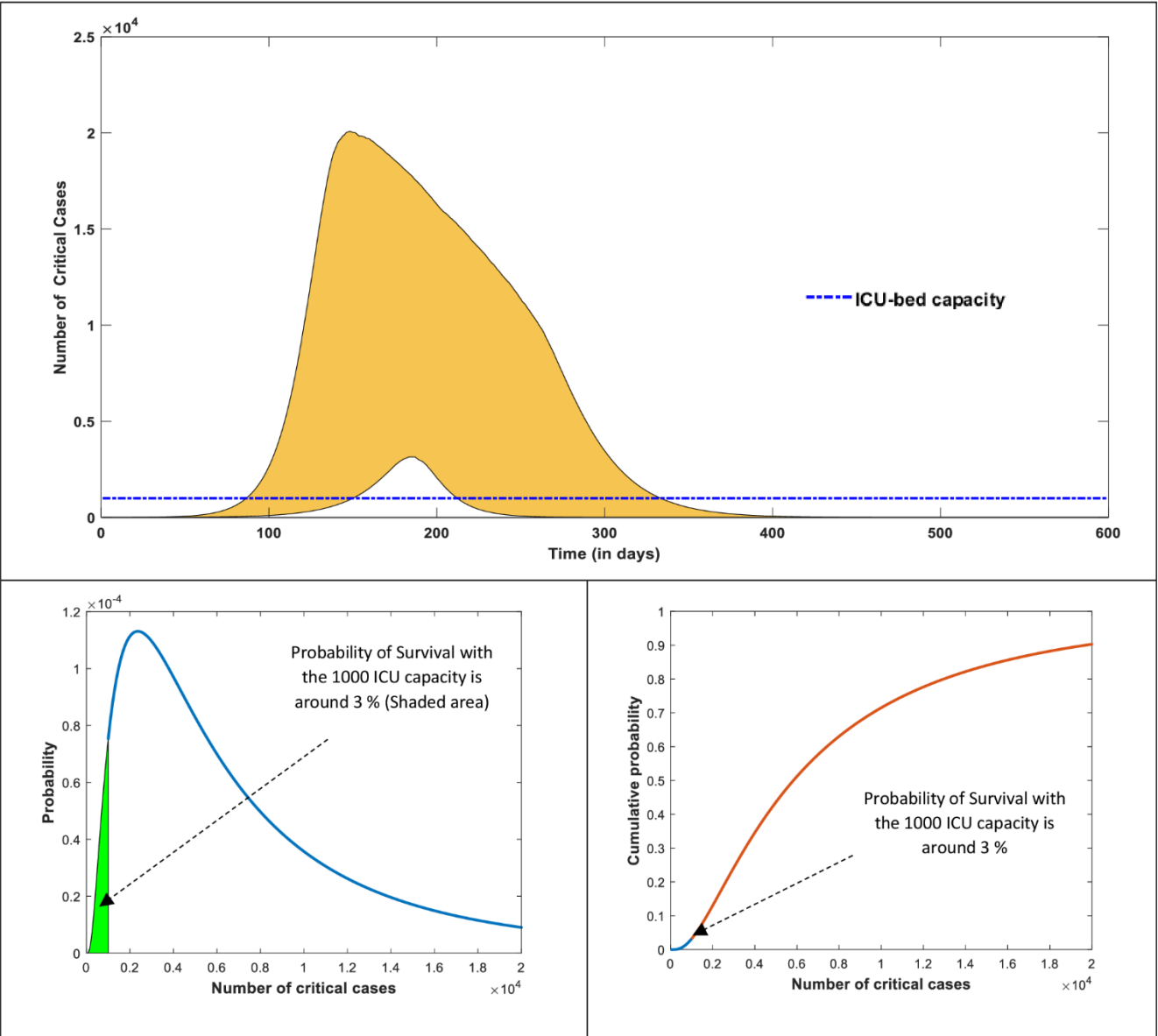
Fig. 8.15: Reliability analysis with the existing healthcare facilities with **no measures** enforced to restrict the COVID-19 pandemic.
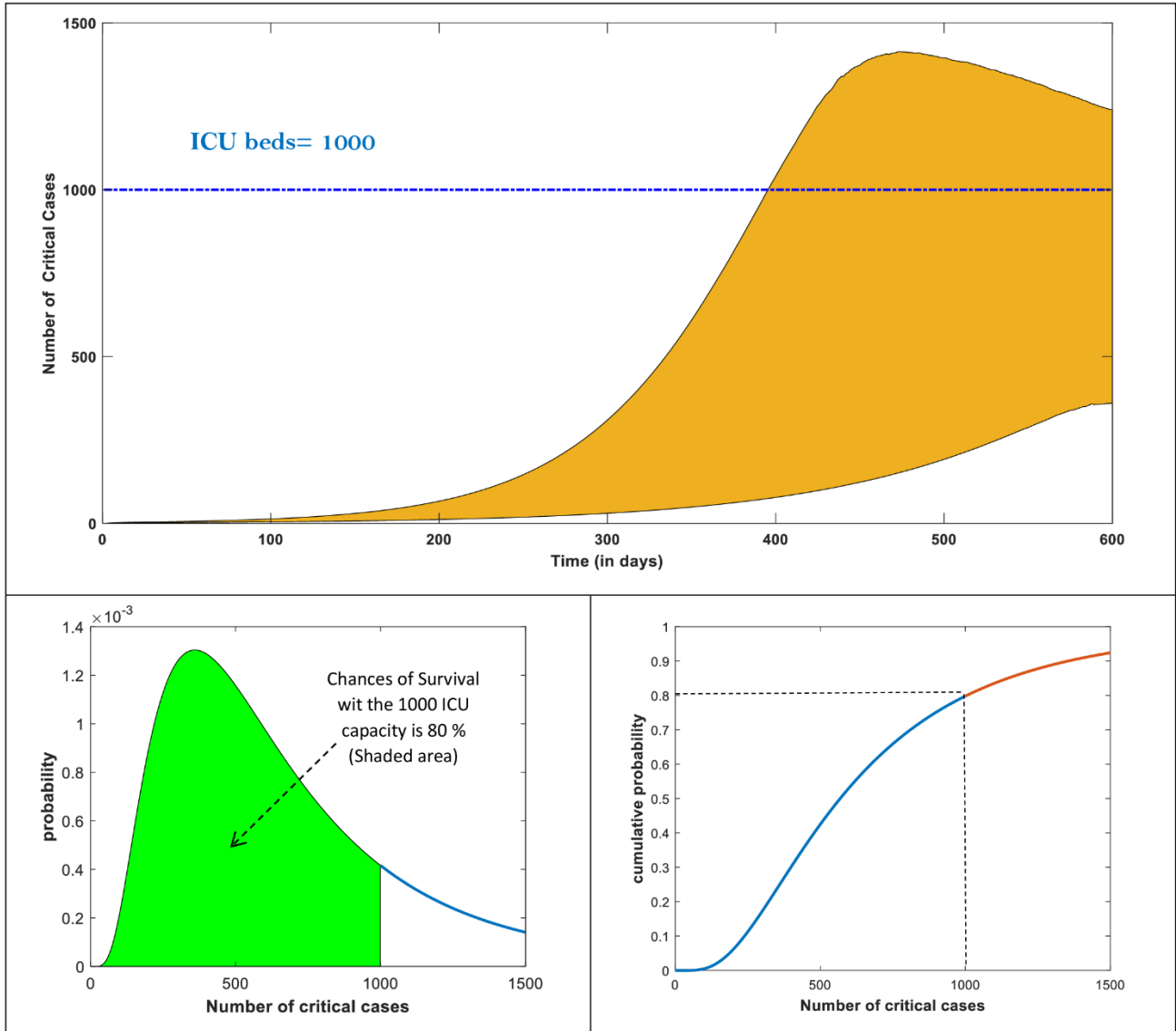
Fig. 8.16: Reliability analysis with the existing healthcare facilities with **School/business closures** enforced to
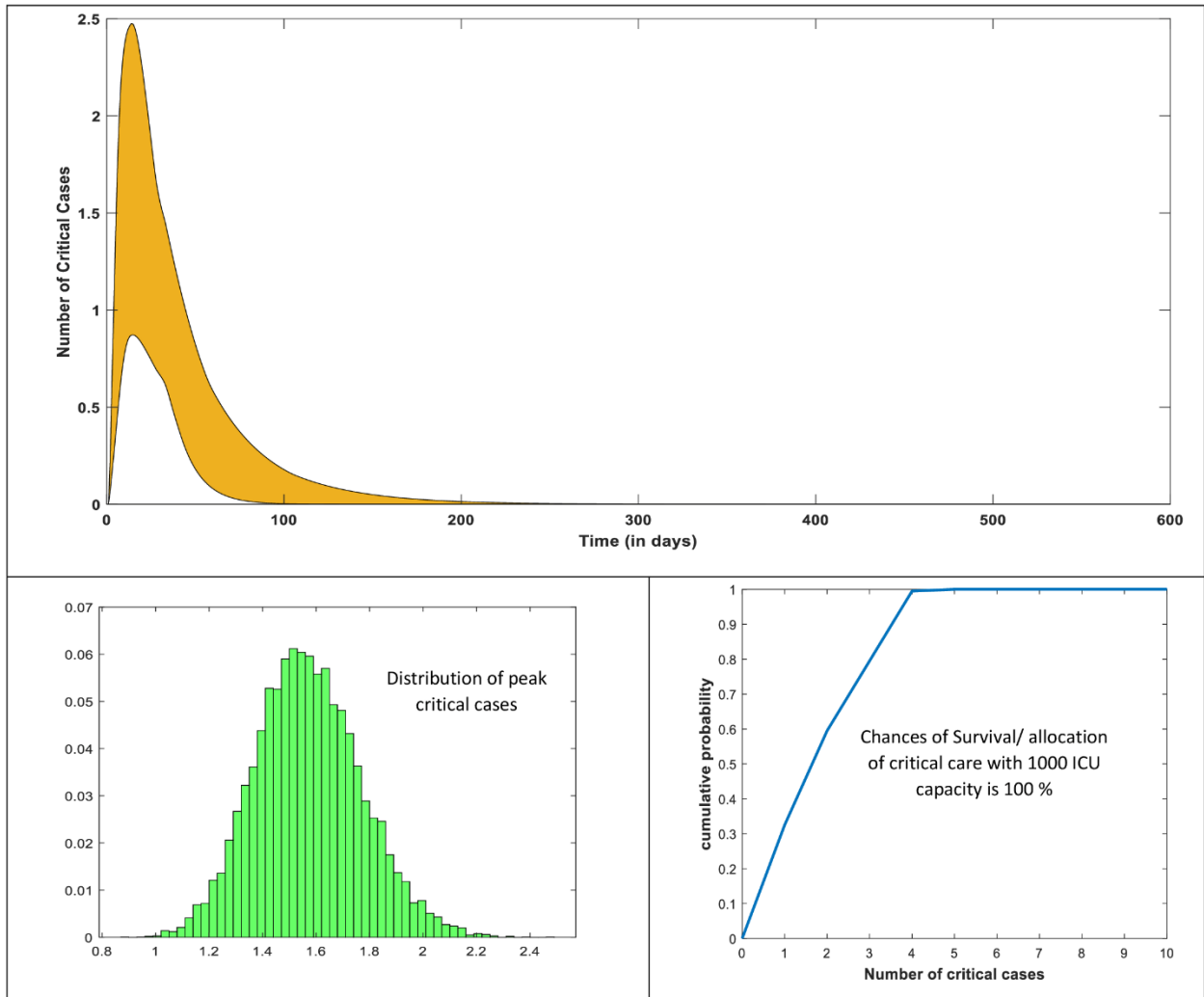
Fig. 8.17: Reliability analysis with the existing healthcare facilities with lockdown to restrict COVID-19.

8.4 Conclusions

This work explores the risk management of a pandemic using engineering safety approaches. The pandemic risk management approaches have been categorized into distinct hierarchical risk reduction strategies: inherent, active, passive, and procedural. We have highlighted how passive control strategies can help mitigate the present and future infectious diseases risk. The impact of the epidemic on an infected individual and the community under distinct scenarios was outlined. We have also developed an event tree diagram for pandemic risk management under assorted barriers such as natural evolution, government interventions, societal responses, and individual practices. Finally, the survival of an infected individual with existing healthcare systems has been investigated for different intervention strategies.

The risk analysis in terms of the number of infections and mortality was performed using *precautionary* and *as low as reasonably practicable* principles. We have included the notion of probability to account for the imapact of the disease using Pate-Cornel's six levels of analysis. The risk calculations were carried out using a the SEIQRD model along with the Monte Carlo simulations. The results show that the implementation of non-pharmaceutical interventions has a profound effect on reducing the risk. The case study demonstrated that the PP and ALARP are applicable in the pandemic-containment decision-making process.

This work does not take into account other fatalities arising from the interruption in health services for chronic disease. Many surveys highlighted the partial or complete disruption of healthcare for hypertension, diabetes-related complications, cancer treatment, and cardiovascular emergencies due to imposed regulations in the COVID-19 pandemic. Moreover, the present work does not capture the vulnerability factor in the analyses, which could be addressed in future works. The model could also be improved by dividing populations based on demographics, spatial dispersion,

and interaction patterns. Rapid testing, contact tracing, and isolation which are critical to controlling disease transmission can also be incorporated for potential improvement.

The test models illustrate the effectiveness of distinct strategies in containing a pandemic with minimal fatality. Lockdown was the most effective measure for reducing risk, but we have no credible estimate of how much reduction came from voluntary isolation and social distancing. This supports many analyst's claims of saving lives using lockdowns. However, we are not advocating for the strict lockdown as its devastating impacts on the economy and mental health cannot be undermined. The stringent lockdown and prolonged confinement can cause neuropsychiatric problems, psychological disorders, and weakened immune systems. A holistic approach with strong ethical and sensible measures is required for combating the epidemic spread (Institute of Medicine, 2007). We have to be prompt in all facets of the transmission; adequate testing facilities, active surveillance, enforcing intervention strategies, and community screening around the cluster areas. Many researchers advocated for "smarter lockdowns" based on granular epidemiological data, temporal segregation, and the social bubble model that allows interaction within a defined group of people while adhering to physical distancing rules with those outside that group (Dhillon & Karan, 2020; Greg, 2020). The extensive support and public endorsement can be asserted by effectively communicating the preparedness and response strategies. The migration and other cross-border entries pose the risk of further spreading an outbreak; it must be handled effectively (Mowat & Raafi, 2020; WHO, 2018).

The real risk of a pandemic is difficult to assess due to uncertainty in several aspects such as the mechanism of COVID-19 transmission, uncertainty in the R-value, randomness in incubation, infection, and recovery periods. Nonetheless, the risk can be alleviated by adopting evidence-based holistic approaches with clear ethical and rational measures such as adequate testing facilities,

active surveillance, enforcing intervention strategies, community screening around the cluster areas.

Acknowledgment

List of symbols and abbreviations

Symbols	*Meanings*
a	contagion rate
ALARP	as low as reasonably practicable
BN	Bayesian network
BTA	bow-tie analysis
c	infection rate
CCU	critical care unit
COVID-19	Coronavirus disease of 2019
D, D(t)	deceased
e	recovery rate
E, E(t)	exposed
ETA	event tree analysis
FMEA	failure mode and effects analysis
FTA	fault tree analysis
HAZOP	hazard and operability study
I, I(t)	infected
ICU	intensive care unit
ISD	inherently safer design
LOPA	layer of protection analysis
MERS	Middle East respiratory syndrome
N, N(t)	total population
NPIs	non-pharmaceutical interventions
PPE	personal protective equipment
PRA	probabilistic risk analysis
Q, Q(t)	quarantined
R, R(t)	recovered
R0	basic reproduction number

S, S(t)	susceptible
SARS	severe acute respiratory syndrome
SARS-CoV-2	severe acute respiratory syndrome coronavirus-2
SEIQRD	susceptible, exposed, infected, quarantined, recovered, deceased
SEIR	susceptible, exposed, infected, and recovered
SIR	susceptible, infected, recovered
T1	infection period
T2	infection period
T3	recovery period
WHO	world health organization
φ_2	fraction of the quarantined/hospitalized population resulting in mortality
φ_1	fraction of symptomatic infections

References

Abate, S. M., Ali, S. A., Mantfardo, B., & Basu, B. (2020). Rate of intensive care unit admission and outcomes among patients with coronavirus: A systematic review and Meta-analysis. *PLoS ONE*, *15*(7 July), e0235653. https://doi.org/10.1371/journal.pone.0235653

Alauddin, M., Islam Khan, M. A., Khan, F., Imtiaz, S., Ahmed, S., & Amyotte, P. (2020). How can process safety and a risk management approach guide pandemic risk management? *Journal of Loss Prevention in the Process Industries*, *68*, 104310. https://doi.org/10.1016/j.jlp.2020.104310

Alvarado, A., Vedantam, S., Goethals, P., & Nopens, I. (2012). A compartmental model to describe hydraulics in a full-scale waste stabilization pond. *Water Research*, *46*(2), 521–530. https://doi.org/10.1016/j.watres.2011.11.038

Amyotte, P., Irvine, Y., & Khan, F. (2018). Chemical safety board investigation reports and the hierarchy of controls: Round 2. *Process Safety Progress*, *37*(4), 459–466. https://doi.org/10.1002/prs.12009

Anderson, R. M., & May, R. M. (1979). Population biology of infectious diseases: Part I. *Nature*, *280*(5721), 361–367. https://doi.org/10.1038/280361a0

Aven, T. (2016). Risk assessment and risk management: Review of recent advances on their foundation. *European Journal of Operational Research*, *253*(1), 1–13. https://doi.org/10.1016/j.ejor.2015.12.023

Baker, T., Schell, C. O., Petersen, D. B., Sawe, H., Khalid, K., Mndolo, S., Rylance, J., McAuley, D., F., Roy, N., Marshall, J., Wallis, L., Molyneux, E. (2020). Essential care of critical illness must not be forgotten in the COVID-19 pandemic. *The Lancet*, *395*(10232), 1253–1254. https://doi.org/10.1016/S0140-6736(20)30793-5

Barrett, K., Khan, Y. A., MBiotech, S. Mac, Ximenes, R., Naimark, D. M. J., & Sander, B. (2020). Estimation of covid-19induced depletion of hospital resources in Ontario, Canada. *Cmaj*, *192*(24), E640–E646. https://doi.org/10.1503/cmaj.200715

BBC. (2020). White House hosted Covid "superspreader" event, says Dr Fauci - BBC News. Retrieved October 24, 2020, from https://www.bbc.com/news/election-us-2020-54487154

Bermingham, S. K., Kramer, H. J. M., & Van Rosmalen, G. M. (1998). Towards on-scale crystalliser design using compartmental models. *Computers and Chemical Engineering*, *22*(SUPPL.1), S355–S362. https://doi.org/10.1016/s0098-1354(98)00075-1

Bloomberg. (2021). More Than 486 Million Shots Given: Covid-19 Tracker, Retrieved March 24, 2021, from https://www.bloomberg.com/graphics/covid-vaccine-tracker-global-distribution/

Cameron, I., Mannan, S., Németh, E., Park, S., Pasman, H., Rogers, W., & Seligmann, B. (2017). Process hazard analysis, hazard identification and scenario definition: Are the conventional tools sufficient, or should and can we do much better? *Process Safety and Environmental Protection*, *110*, 53–70. https://doi.org/10.1016/j.psep.2017.01.025

Chowell, G. (2017). Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. *Infectious Disease Modelling*, *2*(3), 379–398. https://doi.org/10.1016/j.idm.2017.08.001

Courage, K. H. (2020). Caul's cluster | CBC News. Retrieved October 24, 2020, from https://newsinteractives.cbc.ca/longform/cauls-cluster

Cox, L. A. (2012). Confronting Deep Uncertainties in Risk Analysis. *Risk Analysis*, *32*(10), 1607–1629. https://doi.org/10.1111/j.1539-6924.2012.01792.x

Crowl, D. A., & Louvar, J. F. (2011). 12.3-Fault trees. In *Chemical Process Safety Fundamentals with Applications, 3rd edition*.

Cui, L., Zhao, J., & Zhang, R. (2010). The integration of HAZOP expert system and piping and instrumentation diagrams. *Process Safety and Environmental Protection*, *88*(5), 327–334. https://doi.org/10.1016/j.psep.2010.04.002

Cui, Y. Q., Van Der Lans, R. G. J. M., Noorman, H. J., & Luyben, K. C. A. M. (1996).

Compartment mixing model for stirred reactors with multiple impellers. *Chemical Engineering Research and Design*, *74*(2), 261–271.

Danmeng M., & Jia, D. (2020). Lessons From Hong Kong's Covid-19 Superspreading Events. Retrieved October 24, 2020, from https://www.caixinglobal.com/2020-09-23/lessons-from-hong-kongs-covid-19-superspreading-events-101608416.html

Davies, N. G., Kucharski, A. J., Eggo, R. M., Gimma, A., Edmunds, W. J., (2020). Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. *The Lancet Public Health*, *5*(7), e375–e385. https://doi.org/10.1016/S2468-2667(20)30133-X

Dhillon R., S., & Karan, A., 2020. The U.S. Needs Smarter Lockdowns. Now. *Harvard Business Review. https://hbr.org/2020/08/the-u-s-needs-smarter-lockdowns-now*

Dowell, A. M. (1999). Layer of Protection Analysis and Inherently Safer Processes. *Process Safety Progress*, *18*(4), 214–220. https://doi.org/10.1002/prs.680180409

Ferguson, N. M., Cummings, D. A. T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., Burke, D. S. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia, *Nature, 437*(September), 209–214. https://doi.org/10.1038/nature04017

Ferguson, N. M., Laydon, D., Nedjati-gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., Elsland, S., Thompson, H., Verity, R., Volz, E., Wang, H., Wang, Y., Walker, P.G., Walters, C., Winskill, P., Whittaker, C., Donnelly, C., A., Riley, S., Ghani, A. C. (2020). Impact of non-pharmaceutical interventions (NPIs) to reduce COVID- 19 mortality and healthcare demand, London: Imperial College London, March 16.

Frieden, T. R., & Lee, C. T. (2020). Identifying and interrupting superspreading events-implications for control of severe acute respiratory syndrome coronavirus 2. *Emerging Infectious Diseases*, *26*(6), 1061–1066. https://doi.org/10.3201/EID2606.200495

Goerlandt, Floris, & Reniers, G. (2016). On the assessment of uncertainty in risk diagrams. *Safety Science*, *84*, 67–77. https://doi.org/10.1016/j.ssci.2015.12.001

Gollier, C., & Treich, N. (2003). Decision-Making under Scientific Uncertainty: The Economics of the Precautionary Principle. *Journal of Risk and Uncertainty*, *27*(1), 77–103.

https://doi.org/10.1023/A:1025576823096

Giordano, G., Blanchini, F., Bruno, R., Colaneri, P., Di Filippo, A., Di Matteo, A., & Colaneri, (2020). Modelling the COVID-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, *26*(6), 855–860. https://doi.org/10.1038/s41591-020-0883-7

Goyal, A., Reeves, D. B., Cardozo-Ojeda, E. F., Schiffer, J. T., & Mayer, B. T. (2020). Wrong person, place and time: Viral load and contact network structure predict SARS-CoV-2 transmission and super-spreading events. *MedRxiv*, 2020.08.07.20169920. https://doi.org/10.1101/2020.08.07.20169920

Grasselli, G., Pesenti, A., & Cecconi, M. (2020). Critical Care Utilization for the COVID-19 Outbreak in Lombardy, Italy: Early Experience and Forecast during an Emergency Response. *Journal of the American Medical Association*, 323(16), 1545-1546. https://doi.org/10.1001/jama.2020.4031

Greg Ip, (2020, Aug 24), New Thinking on COVID Lockdowns: They're Overly Blunt and Costly. *The Wall Street Journal*. https://www.wsj.com/articles/covid-lockdowns-economy-pandemic-recession-business-shutdown-sweden-coronavirus-11598281419

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D., Du, B., Li, L., Zeng, G., Yuen K., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., Li, S., Wang, J., Liang, Z., Peng, Y., Wei, L., Liu, Y., Hu, Y., Peng, P., Wang, J., Liu, J., Chen, Z., Li, G., Zheng, Z., Qiu, S., Luo, J., Ye, C., Zhu, S., and Zhong, N. (2020). Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine*, 382, 1708-1720 https://doi.org/10.1056/nejmoa2002032

Hethcote, H. W. (1976). Qualitative analyses of communicable disease models. *Mathematical Biosciences*, *28*(3–4), 335–356. https://doi.org/10.1016/0025-5564(76)90132-2

Hiorns, R. W., & MacDonald, N. (1982). Time Lags in Biological Models. *Journal of the Royal Statistical Society. Series A (General)*, *145*(1), 140. https://doi.org/10.2307/2981435

Hu, Z., Cui, Q., Han, J., Wang, X., Sha, W. E. I., & Teng, Z. (2020). Evaluation and prediction of the COVID-19 variations at different input population and quarantine strategies, a case study in Guangdong province, China. *International Journal of Infectious Diseases*, *95*, 231–240. https://doi.org/10.1016/j.ijid.2020.04.010

Iliuta, I., Larachi, F., Anfray, J., Dromard, N., & Schweich, D. (2007). Multicomponent multicompartment model for Fischer-Tropsch SCBR. *AIChE Journal*, *53*(8), 2062–2083.

https://doi.org/10.1002/aic.11242

Institute of Medicine (IOM). (2007). *Ethical and Legal Considerations in Mitigating Pandemic Disease. Ethical and Legal Considerations in Mitigating Pandemic Disease*. Washington, D.C.: National Academies Press. https://doi.org/10.17226/11917

Kabyl, A., Yang, M., Abbassi, R., & Li, S. (2020). A risk-based approach to produced water management in offshore oil and gas operations. *Process Safety and Environmental Protection*, *139*, 341–361. https://doi.org/10.1016/j.psep.2020.04.021

Kermack, W.O., McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *115*(772), 700–721. https://doi.org/10.1098/rspa.1927.0118

Khakzad, N., Khan, F., & Amyotte, P. (2013). Dynamic safety analysis of process systems by mapping bow-tie into Bayesian network. *Process Safety and Environmental Protection*, *91*(1–2), 46–53. https://doi.org/10.1016/j.psep.2012.01.005

Khan, F., Rathnayaka, S., & Ahmed, S. (2015). Methods and models in process safety and risk management: Past, present and future. *Process Safety and Environmental Protection*, *98*, 116–147. https://doi.org/10.1016/j.psep.2015.07.005

Launey, Y., Painvin, B., Roquilly, A., Dahyot-Fizelier, C., Lasocki, S., Rousseau, C., Frasca, D., Gacouin, A., Seguin, P. (2020). Factors associated with time to defecate and outcomes in critically ill patients: a prospective, multicentre, observational study. *Anaesthesia*, July 14, anae.15178. https://doi.org/10.1111/anae.15178

Li, M. L., Zhao, Bourdi, H., T., Lami, O., S., Trikalinos, T., A., Trichakis, N., K., Bertisimas, D., (2020). Forecasting COVID-19 and Analyzing the Effect of Government Interventions. medRxiv preprint doi: https://doi.org/10.1101/2020.06.23.20138693

Liu, Y., Gayle, A. A., Wilder-Smith, A., & Rocklöv, J. (2020). The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine*, *27*(2), 1–4. https://doi.org/10.1093/jtm/taaa021

Mowat D. & Raafi, S., (2020). COVID-19: Impacts and Opportunities. https://www.toronto.ca/wp-content/uploads/2020/09/9133-torr-covid19-impacts-opportunities-2020.pdf

Paiva, H. M., Afonso, R. J. M., de Oliveira, I. L., & Garcia, G. F. (2020). A data-driven model to describe and forecast the dynamics of COVID-19 transmission. *PLoS ONE*, *15*(7 July),

e0236386. https://doi.org/10.1371/journal.pone.0236386

Paté-Cornell, M. E. (1996). Uncertainties in risk analysis: Six levels of treatment. *Reliability Engineering and System Safety*, *54*(2–3), 95–111. https://doi.org/10.1016/S0951-8320(96)00067-1

Pike, H., Khan, F., & Amyotte, P. (2020). Precautionary Principle (PP) versus As Low As Reasonably Practicable (ALARP): Which one to use and when. *Process Safety and Environmental Protection*, *137*, 158–168. https://doi.org/10.1016/j.psep.2020.02.026

Renn, O. (1998). The role of risk perception for risk management. *Reliability Engineering and System Safety*, *59*(1), 49–62. https://doi.org/10.1016/S0951-8320(97)00119-1

Sandin, P. (1999). Dimensions of the precautionary principle. *Human and Ecological Risk Assessment (HERA)*, *5*(5), 889–907. https://doi.org/10.1080/10807039991289185

Singh, S., Roy, D., Sinha, K., Parveen, S., Sharma, G., & Joshi, G. (2020). Impact of COVID-19 and lockdown on mental health of children and adolescents: A narrative review with recommendations. *Psychiatry Research*, *293*, 113429. https://doi.org/10.1016/j.psychres.2020.113429

Spiegelhalter, D. J., & Riesch, H. (2011). Don't know, can't know: Embracing deeper uncertainties when analysing risks. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *369*(1956), 4730–4750. https://doi.org/10.1098/rsta.2011.0163

Vrábel, P., Van Der Lans, R. G. J. M., Cui, Y. Q., & Luyben, K. C. A. M. (1999). Compartment model approach: Mixing in large scale aerated reactors with multiple impellers. *Chemical Engineering Research and Design*, *77*(4), 291–302. https://doi.org/10.1205/026387699526223

Vorou, R. M., Papavassiliou, V. G., & Tsiodras, S. (2007). Emerging zoonoses and vector-borne infections affecting humans in Europe. *Epidemiology and Infection*, *135*(8), 1231–1247. https://doi.org/10.1017/S0950268807008527

Weintraub, K. (2020). Boston Biogen event linked to 20K COVID-19 cases: tracing coronavirus. Retrieved October 24, 2020, from https://www.usatoday.com/story/news/2020/08/26/how-superspreader-events-biogen-conference-incubated-coronavirus-research/3437458001/

Williams, D. (2020). How coronavirus spread to 87% of the singers at a Washington choir practice - CNN. Retrieved October 24, 2020, from https://www.cnn.com/2020/05/13/us/coronavirus-

washington-choir-outbreak-trnd/index.html

World Bank. 2020. Global Economic Prospects, June 2020. Washington, DC: World Bank. © World Bank. https://openknowledge.worldbank.org/handle/10986/33748 License: CC BY 3.0 IGO.

World Commission on the Ethics of Scientific Knowledge and Technology. (2005). The Precautionary Principle - UNESCO Digital Library. Retrieved October 24, 2020, from https://unesdoc.unesco.org/ark:/48223/pf0000139578

World Health Organization (2018). Managing epidemics Key facts about major deadly diseases. https://www.who.int/emergencies/diseases/managing-epidemics-interactive.pdf

Xin, S., Zhang, L., Jin, X., & Zhang, Q. (2019). Reconstruction of the fault tree based on accident evolution. *Process Safety and Environmental Protection*, *121*, 307–311. https://doi.org/10.1016/j.psep.2018.11.003

Zhang, C., Wu, J., Hu, X., & Ni, S. (2018). A probabilistic analysis model of oil pipeline accidents based on an integrated Event-Evolution-Bayesian (EEB) model. *Process Safety and Environmental Protection*, *117*, 694–703. https://doi.org/10.1016/j.psep.2018.06.017

Zhao, W., Buffo, A., Alopaeus, V., Han, B., & Louhi-Kultanen, M. (2017). Application of the compartmental model to the gas-liquid precipitation of CO2-Ca(OH)2 aqueous system in a stirred tank. *AIChE Journal*, *63*(1), 378–386. https://doi.org/10.1002/aic.15567

Chapter 9
Conclusions & Recommendations

The objective of the work presented in this thesis is to develop data-driven models for the safety analysis of complex processing systems and human health management. The objective is achieved through three sub-objectives, to develop robust data-driven models for the safety of process systems, to devise semi-mechanistic models for assessing pandemic risk, and to corroborate synergy between process safety and pandemic risk management. The thesis enacted robust models for handling issues related to poor quality data, model generalization, process uncertainty, and random and spurious errors. The robustness has been instilled by exploiting data quality, reconciling data-driven models with mechanistic modeling, integrating meta-learning, and incorporating prior knowledge and expert opinions. From a model development perspective, several innovative formulations such as robust neural network, process dynamics-guided neural network, a hybrid formulation based on metaheuristic approach, and advanced semi-mechanistic models based on artificial neural network and hierarchical Bayesian framework have been devised to address the modeling inadequacies of complex systems. The following sections describe how the thesis accomplished the objective and sub-objectives of the study.

9.1 Contribution to the domain advancement by developing cutting-edge tools

This thesis presents several formulations of data-driven models for operational safety management of complex processing systems.

A robust ANN based on a novel approach, the quality-based training (QbT), has been developed for handling mislabeled and/or low-quality data in a deep neural network. The model imparted improved performance on detecting faults on mislabeled samples. The proposed model yielded

comparatively enhanced outcomes in detecting faults on legitimate data as well. It resulted in up to 5 % improvement in accuracy in detecting faults of test cases comprising legitimate data. The reason for this is the semi-supervised framework of the proposed model that reinforces the supervisory labels using the data quality attributes. This can assist in classifying the boundary cases in normal and faulty classes which could otherwise be intractable by the standard ANN-based exclusively on the supervisory labels. Thus, the proposed robust model is effective in detecting faults in complex processing systems.

The thesis presents a *process dynamics guided neural network (PDNN) model* to advance model generalization by rendering process dynamics and field expertise. The PDNN model aims to reconcile data-driven models with mechanistic wings to develop hybrid models for reliable predictions. The proposed model conferred improved performance on regression and classification tasks of processing systems representing steady-state and transient behavior. The performance was intact on three diverse operations, virtual mixing, separation in a binary distillation and reactant conversion in a batch reactor. The predictive capacity of the proposed PDNN was superior on extrapolated samples and the model trained with relatively fewer data. This is helpful in establishing dependency in a novel process. It can also guide studying processes where the data generation is expensive. Thus, the proposed hybrid model can lead to reliable estimates with enhanced generalization.

The performance of a data-driven method is improved using meta-learning as well. The thesis presents a hybrid model based on artificial neural network (ANN) and variable mosquito flying optimization (V-MFO) technique for detecting faults of complex process systems. The formulation attempts to fine-tune the neural structure by maximizing fault detection rate and minimizing the

false alarm rate. The proposed hybrid model yielded improved results in detecting faults in the Tennessee Eastman benchmark process.

This thesis presents advanced semi-mechanistic models for pandemic risk assessment with a special focus on handling uncertainty, data quality, and the dynamic nature of a system. Here the mechanistic part refers to the science-based structure of the disease dynamics whereas the 'semi' denotes the enhancement instilled using advanced data-driven models. The semi-mechanistic models present efficient frameworks enriched with mechanistic knowledge in the form of a set of differential equations of the disease dynamics, prior knowledge of parameter distribution, and pooling information from distinct periods of the pandemic outbreaks.

This thesis presents an artificial neural network and Bayesian inference-based formalisms to make the assessment more compelling by capturing temporal changes in the parameters of a mechanistic model. Artificial neural networks are data-driven nonlinear modeling techniques with a strong capability to model nonlinear relationships among process variables. A data-driven semi-mechanistic SEIQRD model has been used to develop a risk management framework for effective forecasting of dynamic risk. The effect of enforcement and releasing of the non-pharmaceutical interventions at different stages of the disease outbreak has also been evaluated. Monte Carlo simulation was used to capture the randomness of the model parameters. It facilitated a probable range of predictions.

Bayesian inference is an effective tool for dealing with uncertainty and easier incorporation of prior beliefs. Although a standard Bayesian framework can consider the conditional dependencies among various factors of a complex system, it is not capable of modeling such dependencies under varying conditions. Also, a standard Bayesian network cannot handle data from other similar case studies, the hierarchical Bayesian methods are efficient tools for handling these multisource data.

The thesis explored updating posteriors distribution using multiple sources to enhance the credibility of predictions which could not be possible with a fewer number of representative samples of a particular source. The variability of the SEIQRD parameters in distinct phases of the COVID-19 outbreak has been captured by enabling the parameter sharing feature of the hierarchical Bayesian formalism.

A common practice in Bayesian-based formalisms is to select probability distributions of priors and likelihood functions from conjugate families, e.g., Poisson-Gamma or Beta-Binomial distributions. The conjugate pairs result in a closed-form solution of standard posteriors. However, many real-world data do not follow conjugate behavior and need to be addressed using sophisticated numerical techniques. The thesis presented a hierarchical Bayesian structure with non-conjugate distributions to solve a real-world problem of pandemic risk management. The uncertainty caused due to government interventions (such as lockdown, school and business closures, and restricting large gatherings), changes in local conditions and operational formats (such as enabling home delivery services and working from home), and changes in individual behavior (such as social distancing, wearing a mask, and hygiene practices) have been dealt with a hierarchical Bayesian formalism. The thesis also investigated other Bayesian network formulations based on pooling, partial pooling, and no pooling in handling the temporal variability of a pandemic.

Markov chain Monte Carlo sampling (MCMC) has been employed to sample posterior probabilities from a high-dimensional distribution. Sampling using the MCMC from the actual distribution of the model parameters such as reproduction number, rate of incubation, infection, and recovery is sluggish. Thus, the thesis applied variational inference to approximate the real posterior distribution of parameters. No U-turn sampler (NUTS), an adaptive tuning algorithm

based on a recursive approach, has been used for the MCMC sampling to get effective samples of the approximated probability distributions.

9.2 Establishing interdisciplinary parallelism of process safety principles

This thesis explores the similarity between process safety principles and epidemiology to better understand, analyze, and manage the risk.

The epidemiology and process systems behave identically in mathematical modeling for the risk forecast. Process controllers constantly render control actions to prevent process deviations due to various disturbances, incipient abnormalities, transportation lag, and process dynamics. Equivalently, pandemic risk management addresses the temporal variations caused by imposed regulations, varied individual responses, and the advent of multiple waves of an outbreak. Uncertainty and inconsistency in data are critical factors affecting processing systems and epidemiological modeling. Erroneous and low-quality data frequently materialize in complex processing systems for many reasons, *e.g.* reporting the data from faulty sensors, malfunctioning of equipment, random fluctuations, incorrect calibration, inconsistent sampling frequencies, and human error. Uncertainty and inconsistency are inevitable in epidemiological studies as well. A pandemic modeling comprises numerous sources of uncertainty such as modes of propagation, uncertainty in infectivity, rates of incubation, infection, and recovery periods. Moreover, the reporting of the infected cases is often flawed by numerous factors such as lack of systematic testing, the inherent delay between the date that an illness starts and the date the case is reported to public health authorities.

This thesis discusses similarities between the epidemiological model and chemical processing systems by highlighting the resemblance of the SIR disease dynamics with the reaction kinetics model of a CSTR. The factors governing contagion and reaction rates were underlined. Also,

several areas of similarities have been identified where process safety and epidemiology could benefit from each other. The pandemic risk has been analyzed using process safety principles such as precautionary principles, as low as reasonably practicable (ALARP) approach, event tree model, and the layer of protection analysis. The formulations address the dynamic risk caused by the natural evolution, government interventions, societal responses, and individual practices. This thesis also categorizes distinct risk-reducing strategies into hierarchical safety frameworks.

9.3 Practical relevance and implications of the thesis

Data-driven modeling, represented by machine learning (ML) and artificial intelligence (AI), is at the top of the Gartner hype cycle. Many experts believe that digital transformation is no longer simply a future option for processing systems, it's an urgent reality. But, the data-based models rely on the assumption that the data is of high quality, sufficient granularity, and quantity. They also suffer from the lack of generalization ability and struggle when data is scarce or in extrapolation regimes. The data-based model can also be misleading in presence of low-quality data. A data-driven model based on poor-quality data can result in poor decisions leading to loss of revenue and reputation. Industrial systems encounter various low data-quality problems such as incomplete data, outliers, missing information, imbalanced data, poorly labeled data, incorrectly mapped properties, noise, and inconsistency due to sensor breakdown, process shutdown, malfunctioning of equipment, random fluctuations, incorrect calibration, inconsistent sampling frequencies, and data entry errors due to human factors. Handling low-quality data is crucial in the data-driven models of industry 4.0. Models based on the internet of things (IoT) where data come from multiple sources are susceptible to low quality and inconsistent data. The automated processes are based on sensor data and dealing with many real-time data is challenging. Although most data-based models are preceded by a pretreatment step for relieving missing values and

319

scaling problems, the mislabeled data and inconsistency are unable to be handled using such processing. The proposed robust approach, the quality-based training of data-driven models, can aid to critical operations such as fault detection and diagnosis of process systems which can cause huge losses due to mislabeled or low-quality data.

The lack of interpretability and the black-box nature of the data-driven models are other striking challenges in widespread adoption in industrial systems. First principle models on the other hand rely on system understanding to compensate for lack of data. They suffer from knowledge gaps in describing many frequently occurring processes such as turbulence, random variations, and human factor induced deviations. The thesis presents *process dynamics-guided neural network (PDNN) model* to improve the model generalization by rendering process dynamics and field expertise. Integrating process dynamics is also an attempt to unveil the black-box nature of the deep neural network which is resistant to widespread adoption of these models at an industrial scale. Integrating process dynamics and field knowledge can provide confidence in the data-based models. This thesis also reckoned how the PDNN can lead to robust models with a capacity to explore novelty and the unexplained dynamics of complex processes.

This dissertation also investigated the Coronavirus disease of 2019 that has wreaked havoc worldwide. The thesis attempted to address the startling challenges by devising credible assessment methodologies under the realm of engineering knowledge. The methods can aid decision-making to restrain the current and the future outbreak. This reflects the Wiemen's quote "Education is about learning to make better decisions- Carl Wiemen". Cases for reducing the risk caused due to the COVID-19 pandemic that has inflicted millions of fatalities, afflicted billions of people with food and livelihood, and put existential threats to millions of enterprises are also presented. The pandemic severely hit energy industries leading to a historic collapse in demand

and price of crude and allied products. Not surprisingly, the COVID-19 pandemic also impacted our studies and research work severely. This thesis devised multiple semi-mechanistic models in engineering framework for credible assessment to help 'better decision-making'.

This thesis presented a hierarchical Bayesian framework for administering variabilities caused due to distinct phases of a pandemic outbreak. Although, proposed hierarchical Bayesian method was employed to study the pandemic risk assessment, it can be applied to treat uncertainties of the data associated with multiple sources of complex industrial systems. This can be effective tool for managing the data-quality concerns of the models based on the internet of things (IoT).

9.4 Limitations of the work and recommendations for future research

The present work attempts to introduce new formulations to overcome the limitations of existing data-driven and mechanistic techniques. This thesis presented robust data-driven methods for ensuring safety of public health and processing systems. This study, however, can be extended further to address the limitations of the work. Some of those promising areas are mentioned in the following sections.

a. The thesis presents a robust neural network model by exploiting the data- quality. The study was based on the random mislabeled data in the quality-based robust neural network model. The performance can further be improved if the causality and structural relation is accounted for in the model. Combining physics and domain knowledge with quality-based training can also be studied to analyze robustness of detection systems.

b. Process dynamics guided deep neural network model was developed to improve generalization by rendering process dynamics and field expertise. This work could further be improved by including partial differential equations and complex correlations in the proposed process dynamics layer of the deep neural network. Alarm and other data can

also be used to aid learning of the model parameters. Validation of the developed models with real industrial data is critical. Several packages such as ASPEN Plus, PRO/II, ProSim, CADsim, Chemcad, ChemPro, ChromWorks, and DWSim have been used in process design, monitoring, and control of chemical processing facilities. Integrating the proposed robust data-driven methods in existing simulation packages can accelerate the pace of adoption in industrial systems.

c. This thesis presented a hybrid model for the alarm tuning of the process fault detection systems based on artificial neural network (ANN) and variable mosquito flying optimization (V-MFO) technique. The model yielded improved performance on detecting and isolating faults. However, it does not study its impact on generalization which could have been affected by the exploration of the evolutionary computation. This can also be expanded to accommodate alarm design features for a more holistic improvement.

d. While efforts to use digital solutions in process operations are gaining wider acceptance, there are serious safety concerns that need to be addressed. Cyber-attacks are the key deterrents in the digitization and digitalization of industrial systems. Many industrialists believe that the benefits gain from digitalization is substantial even with increased cybersecurity risk. This can be studied on the prosed semi-mechanistic models of processing systems.

e. This work presents robust models based on integrating mechanistic models and expert knowledge. This, however, did not discuss cases about the conflict between human and data-based models. Incorporating that aspect would be an interesting piece of work for handling real-world scenarios. Noise plays an important role in industrial systems. The

robustness of the models should also be closely investigated for evaluating the impact of the signal-noise interaction on data acquisition systems.

f. This thesis presents the SEIQRD model that can be made comprehensive by including testing, contact tracing, isolation, and vaccination which are critical to controlling disease transmission. This study assumes a blunt lockdown that is strictly being followed throughout a geographical region. Many researchers advocated for "smarter lockdowns" based on granular epidemiological data, temporal segregation, and the social bubble model that allows interaction within a defined group of people while adhering to physical distancing rules with those outside that group. The assessment can be improved based on granular data. This work does not consider other fatalities arising from the interruption in health services for chronic disease. Many surveys highlighted the partial or complete disruption of healthcare for hypertension, diabetes-related complications, cancer treatment, and cardiovascular emergencies due to imposed regulations in the COVID-19 pandemic. Moreover, the present work does not capture the vulnerability factor in the analyses which can be a suitable topic to be addressed in future works.

g. This thesis employed several process safety principles for pandemic risk management. Although pointed out symbolically, however, it did not provide any explicit formulation for enhancing process safety using pandemic risk management approaches. Several responses to combating disease progression could be harnessed to improve the safety of process systems. For example, the working of the human immune system that acts as the first layer of defense against disease progression could be imitated for enhancing the resilience of chemical processing systems.

Appendices

Appendix A: Basic reproduction numbers from distinct studies (adapted from (Liu, Gayle, Wilder-Smith, & Rocklöv, 2020))

Study	Location Study	Study date Methods	Methods Approaches	Approaches	R0 estimates (average)	95% CI
Joseph et al.1	Wuhan	31 December 2019–28 January 2020	Stochastic Markov Chain Monte Carlo methods (MCMC)	MCMC methods with Gibbs sampling and non-informative flat prior, using posterior distribution	2.68	2.47–2.86
Shen et al.2	Hubei province	12–22 January 2020	Mathematical model, dynamic compartmental model with population divided into five compartments: susceptible individuals, asymptomatic individuals during the incubation period, infectious individuals with symptoms, isolated individuals with treatment and recovered individuals	R0 = β/α β = mean person-to-person transmission rate/day in the absence of control interventions, using nonlinear least squares method to get its point estimate α = isolation rate = 6	6.49	6.31–6.66

Liu et al	China and overseas	23 January 2020 Statistical	Statistical exponential Growth, using SARS generation time=8.4 days, SD=3.8 days	Applies Poisson regression to fit the exponential growth rate R0 =1/M(−r) M=moment generating function of the generation time distribution r=fitted exponential growth rate	2.90	2.32-3.63
Liu et al	China and overseas	23 January 2020 Statistical	Statistical maximum likelihood estimation, using SARS generation time=8.4 days, SD=3.8 days	Maximize log-likelihood to estimate R0 by using surveillance data during a disease epidemic, and assuming the secondary case is Poisson distribution with expected value R0	2.92	2.28-3.67
Read et al	China	1–22 January 2020	Mathematical transmission model assuming latent period=4 days and near to the incubation period	Assumes daily time increments with Poisson-distribution and apply a deterministic SEIR metapopulation transmission model, transmission rate=1.94, infectious period =1.61 days	3.11	2.39–4.13
Majumder et al	Wuhan	8 Dec, 2019 and 26 January 2020	Mathematical Incidence Decay and Exponential Adjustment (IDEA) model	Adopted mean serial interval lengths from SARS and MERS ranging from 6 to 10 days to fit the IDEA model,	2.55	2.0–3.1

					1.95	1.4–2.5
WHO	China	18 January 2020				
Cao et al	China	23 January 2020	Mathematical model including compartments Susceptible-Exposed-Infectious-Recovered-Death-Cumulative (SEIRDC)	R = K 2(L × D) + K(L + D)+1 L=average latent period=7, D=average latent infectious period=9, K=logarithmic growth rate of the case count	4.08	
Zao et al	China	10–24 January 2020	Statistical exponential growth model method adopting serial interval from SARS (mean=8.4 days, SD=3.8 days) and MERS (mean=7.6 days, SD=3.4 days)	Corresponding to 8-fold increase in the reporting rate R0 =1/M(−r) r =intrinsic growth rate M=moment generating function	2.24	1.96–2.55
Zhao et al	China	10–24 January 2020	Statistical exponential growth model method adopting serial interval from SARS (mean=8.4 days, SD=3.8 days) and MERS (mean=7.6 days, SD=3.4 days)	Corresponding to 2-fold increase in the reporting rate R0 =1/M(−r) r =intrinsic growth rate M=moment generating function	3.58	2.89-4.39
Imai (2020)	Wuhan	January 18, 2020	Mathematical model, computational modelling of potential epidemic trajectories	Assume SARS-like levels of case-to-case variability in the numbers of secondary cases and a SARS-like generation time with 8.4 days, and	2.5	1.5–3.5

					set number of cases caused by zoonotic exposure and assumed total number of cases to estimate R0 values for best-case, median and worst-case		
Julien and Althaus	China and overseas	18 January 2020	Stochastic simulations of early outbreak trajectories Tang	Stochastic simulations of early outbreak trajectories were performed that are consistent with the epidemiological findings to date	2.2	
Tang et al	China	22 January 2020	Mathematical SEIR-type epidemiological model incorporates appropriate compartments corresponding to interventions	Method-based method and Likelihood-based method	6.47	5.71–7.23
Qun Li et al.11	China	22 January 2020	Statistical exponential growth model	Mean incubation period=5.2 days, mean serial interval=7.5 days	2.2	1.4–3.9
Steven et al	China (CDC)			Realistic distributions for the latent and infectious period to calculate R_0	5.7	3.8-8.9

Avarage R_0= 3.4　　　　　　　Median R_0= 2.9

Appendix B: Derivation of the basic reproduction number for the SEIQRD model using the next generation method

Let X represent the seven states of the SEIQRD model using x1, x2, x3, x4, x5, x6, and x7.

$$F_1(X) = \begin{bmatrix} \frac{a}{N}x_1x_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, V1(X) = \begin{bmatrix} bx_2 \\ -bx_2 + cx_3 \\ -\varphi_1cx_3 + dx_4 \\ -dx_4 + ex_5 \end{bmatrix}$$

$$F = \frac{dF_1(X)}{dx_1} = \begin{bmatrix} 0 & a & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad V = \frac{dV1(X)}{dx_1} = \begin{bmatrix} b & 0 & 0 & 0 \\ -b & c & 0 & 0 \\ 0 & -\varphi_1c & d & 0 \\ 0 & 0 & -d & e \end{bmatrix},$$

The inverse of $\frac{dV1(X)}{dx_1}$ would be obtained using the row operations as-

$$\begin{bmatrix} b & a & 0 & 0 \\ -b & c & 0 & 0 \\ 0 & -\varphi_1c & d & 0 \\ 0 & 0 & -d & e \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xrightarrow{R_2 \to R_1 + R_2} \begin{bmatrix} b & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & -\varphi_1c & d & 0 \\ 0 & 0 & -d & e \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\xrightarrow{R_3 \to \varphi_1 R_2 + R_3} \begin{bmatrix} b & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & d & 0 \\ 0 & 0 & -d & e \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \varphi_1 & \varphi_1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \xrightarrow{R_4 \to R_3 + R_4} \begin{bmatrix} b & 0 & 0 & 0 \\ 0 & c & 0 & 0 \\ 0 & 0 & d & 0 \\ 0 & 0 & 0 & e \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \varphi_1 & \varphi_1 & 1 & 0 \\ \varphi_1 & \varphi_1 & 1 & 1 \end{bmatrix}$$

$$
\begin{array}{l}
R_1 \rightarrow \dfrac{R_1}{b} \\[2mm]
R_2 \rightarrow \dfrac{R_2}{c} \\[2mm]
R_3 \rightarrow \dfrac{R_3}{d} \\[2mm]
R_4 \rightarrow \dfrac{R_4}{e}
\end{array}
\rightarrow
\begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\dfrac{1}{b} & 0 & 0 & 0 \\[2mm]
\dfrac{1}{c} & \dfrac{1}{c} & 0 & 0 \\[2mm]
\dfrac{\varphi_1}{d} & \dfrac{1}{d} & 1 & 0 \\[2mm]
\dfrac{\varphi_1}{d} & \dfrac{\varphi_1}{e} & \dfrac{1}{e} & \dfrac{1}{e} \\[2mm]
\dfrac{\varphi_1}{e} & & &
\end{bmatrix}
$$

$$
FV^{-1} =
\begin{bmatrix}
0 & a & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
\dfrac{1}{b} & 0 & 0 & 0 \\[2mm]
\dfrac{1}{c} & \dfrac{1}{c} & 0 & 0 \\[2mm]
\dfrac{\varphi_1}{d} & \dfrac{1}{d} & 1 & 0 \\[2mm]
\dfrac{\varphi_1}{d} & \dfrac{\varphi_1}{e} & \dfrac{1}{e} & \dfrac{1}{e} \\[2mm]
\dfrac{\varphi_1}{e} & & &
\end{bmatrix}
=
\begin{bmatrix}
0 & \dfrac{a}{c} & \dfrac{a}{c} & 0 \\[2mm]
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix}
$$

The spectral radius of FV^{-1} could be-
$$
\left| \begin{bmatrix}
0 & \dfrac{a}{c} & \dfrac{a}{c} & 0 \\[2mm]
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0
\end{bmatrix} - \lambda I \right| = 0
$$

$$
\lambda^2 \left(\frac{a}{c} - \lambda \right) = 0 \rightarrow \quad \lambda = 0, \qquad \lambda = \frac{a}{c}
$$

$$
\lambda = 0 \; is \; not \; relevent \; for \; the \; epidemic \; spread,
$$

$$
Thus, \; the \; reproduction \; number \, , \;\; R_0 = \frac{a}{c}
$$

A simple posterior plot can be created using the traceplot. The left column consists of a smoothed histogram (using kernel density estimation) of the marginal posteriors of each stochastic random variable while the right column contains samples of the Markov chain plotted in sequential order.

R0_1

R0_2

T0_1

T0_2

T1_1
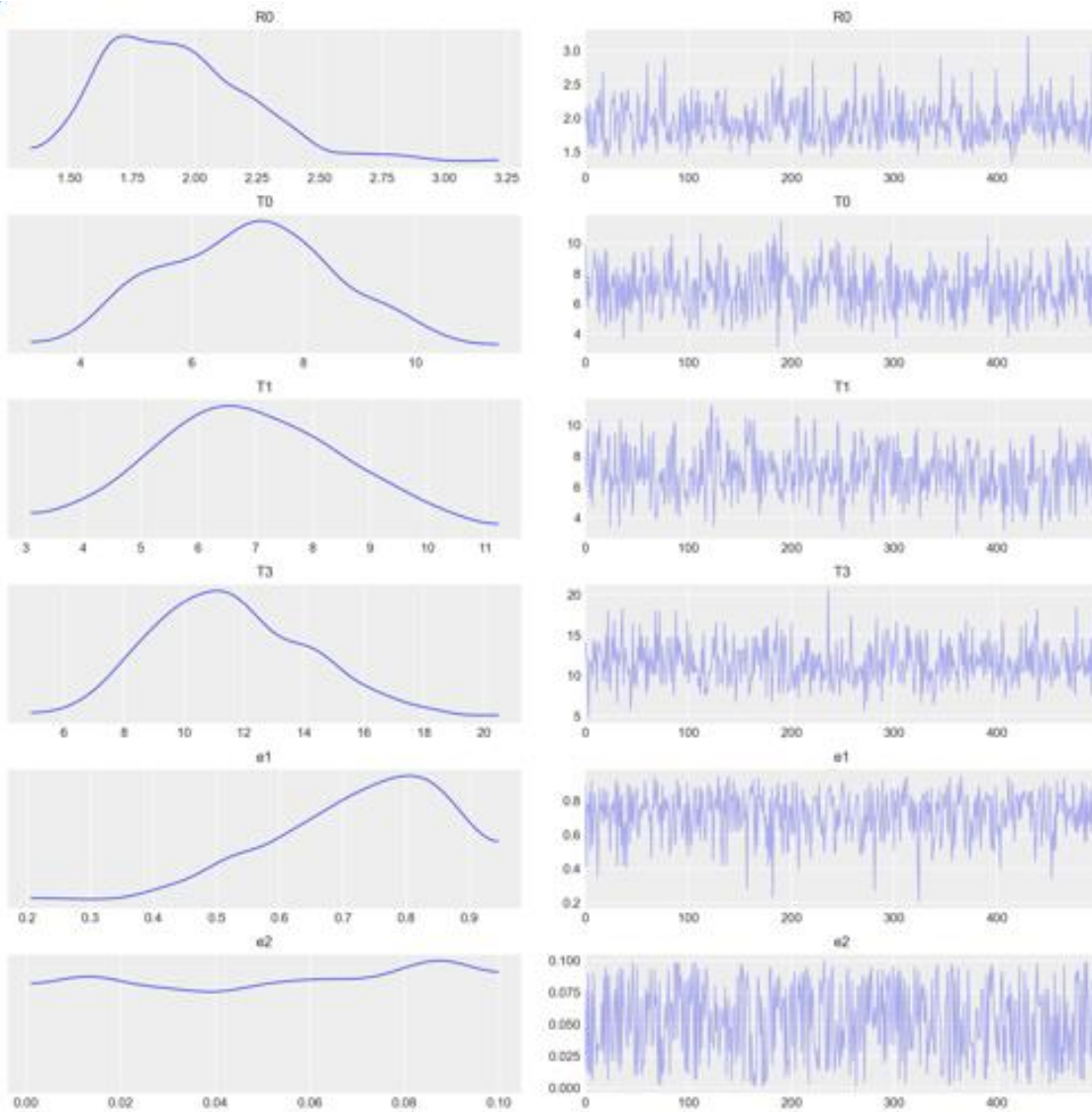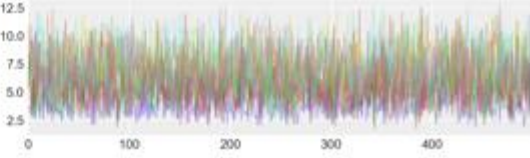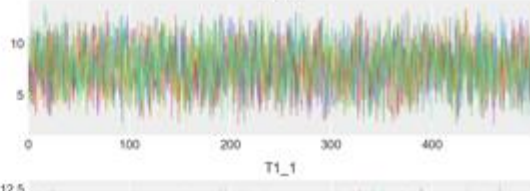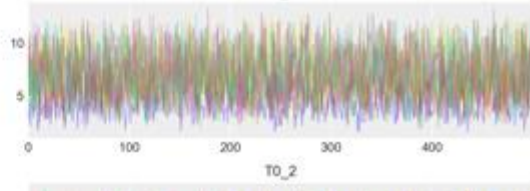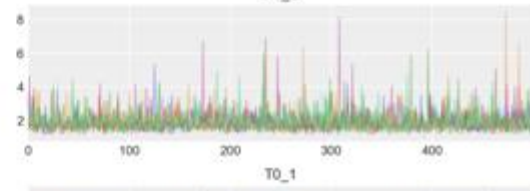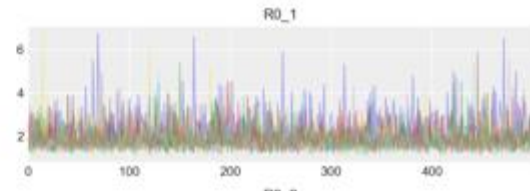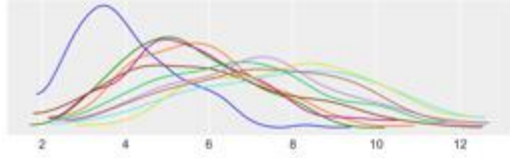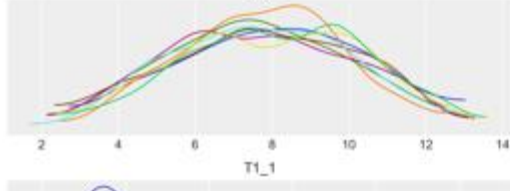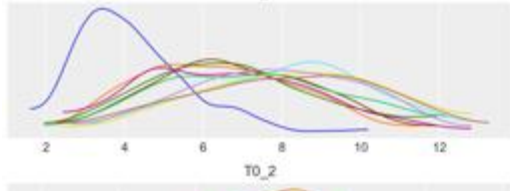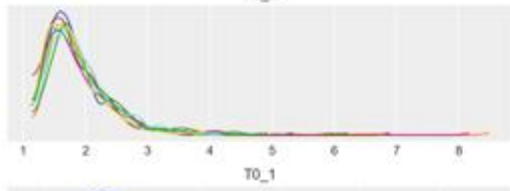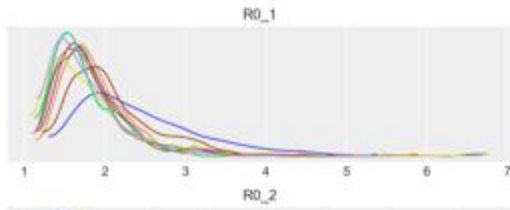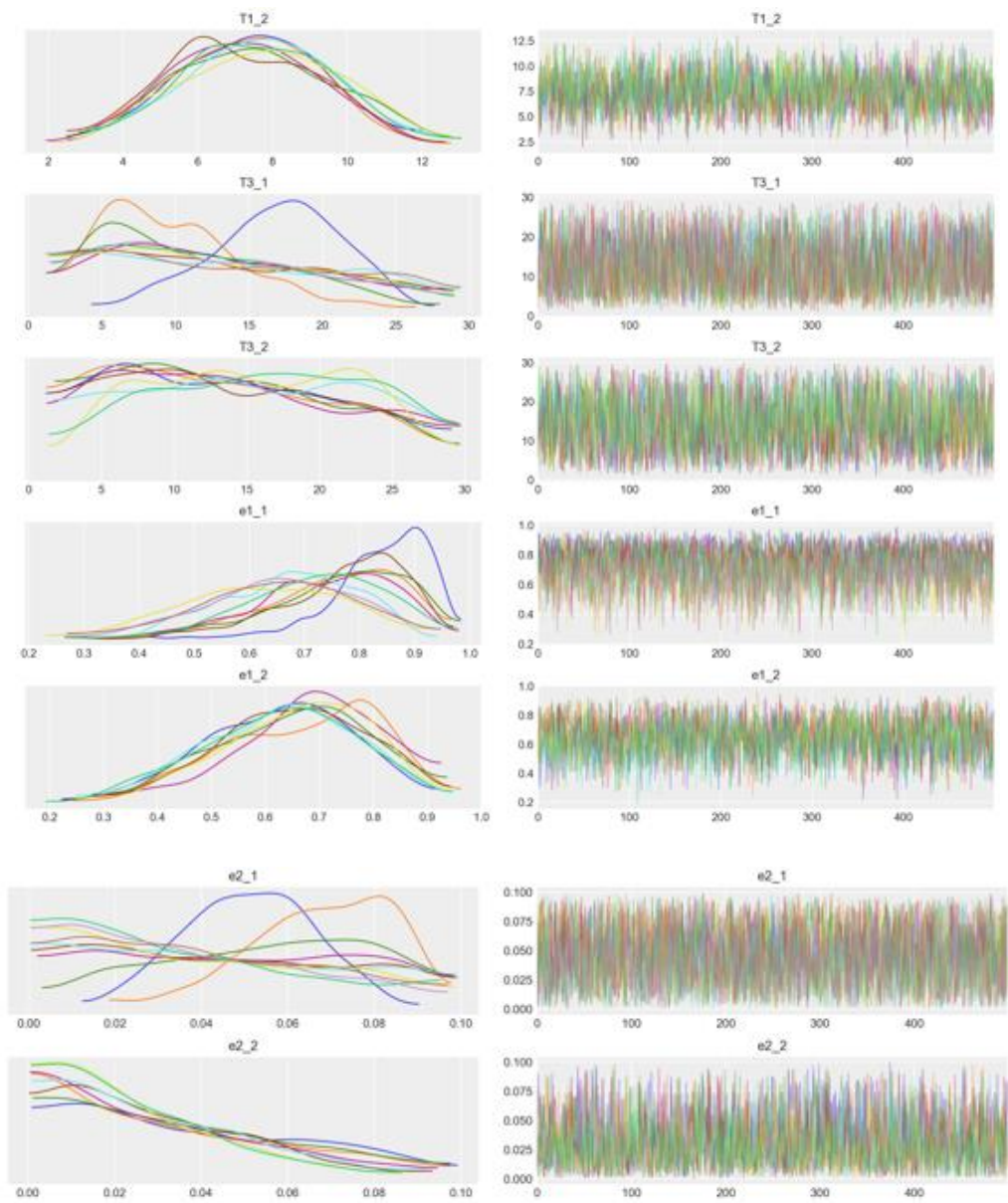
```python
#importing libraries
import arviz as az
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import pymc3 as pm
import theano as tt
import theano
from pymc3.ode import DifferentialEquation
from scipy.integrate import odeint
np.random.seed(42)
pm.set_tt_rng(42)
plt.style.use("seaborn-darkgrid")

# Load the data
Data=pd.read_csv("Suppl_S2_ Ontario_2.csv")
yobs=data.values

Data1=pd.read_csv("Suppl_S1_ Ontario_Mar_2020_to_Jan_2021.csv")
yobss=data1.values
# Data preprocessing
from scipy.signal import savgol_filter
tp=np.arange(1, 300, 1)
#I2_savgol0  = savgol_filter(yobss[:,5], 15, 0) # window size 51, polynomial
order 1
#I2_savgol1  = savgol_filter(I2_savgol0[:], 15, 1) # window size 51,
polynomial order 1
yobs=yobss.copy();
#yobs=copy.deepcopy(yobss);
#copy.deepcopy
yobs[:,3]=I2_savgol1[:]

import seaborn as sns
fig, ax=plt.subplots()
ax.plot(tp1,yobss[tp1,4],color='g', label='raw data',linewidth=2.0)
ax.plot(tp1, yobs[tp1,4],color='r',label='smooth data',linewidth=2.0)
sns.despine()
ax.legend()

## Defining the function for the SEIQRD model
def SEIQRD(y, t, p):
    N=14600000                  # Population of the Location
    ds = -p[0] * y[0] * y[2]/N
    de = p[0] * y[0] * y[2]/N - p[1] * y[1]
    di1 = p[1] * y[1]-p[2]*y[2]
    di2 = p[2]*p[4]*y[2]-y[3]
```

```python
    dq =  y[3]-p[3]*y[4]
   # dr = p[2]*(1-p[4])*y[2]+(1-p[5])*p[3]*y[4]
    dd = p[5]*p[3]*y[4]
    return [ds, de, di1, di2, dq, dd]

model_SEIQRD = DifferentialEquation(
    func=SEIQRD,
    times=np.arange(1, 16, 1),
    n_states=6,
    n_theta=6,
    t0=0,
)

##Function for generating the SEIQRD based data for Hierarchical and unPooled
Bayesian models
def mymodel_B66(R0_1,R0_2, T0_1,T0_2, T1_1,T1_2, T3_1,T3_2, e1_1,e1_2,
e2_1,e2_2, idx):
    N=14600000
    s1=[]
    allArrays=[]
    j=0
    for i in range (int(idx)):
            if i<=9:         # for the first wave
                  RR=R0_1[i]/T1_1[i]
                  bb=1/T0_1[i]
                  cc=1/T1_1[i]
                  dd=1/T3_1[i]
                  ee1=e1_1[i]
                  ee2=e2_1[i]
                  ss = model_SEIQRD(y0=yobs[15*i], theta=[RR, bb, cc, dd,
ee1, ee2])
                  s1.append(ss)

            else:               # for the second wave
                  RR=R0_2[j]/T1_2[j]
                  bb=1/T0_2[j]
                  cc=1/T1_2[j]
                  dd=1/T3_2[j]
                  ee1=e1_2[j]
                  ee2=e2_2[j]
                  ss = model_SEIQRD(y0=yobs[15*i], theta=[RR, bb, cc, dd,
ee1, ee2])
                  s1.append(ss)
            j=j+1
      sss13 = theano.tensor.concatenate([s[0], s[1], s[2], s[3], s[4],
    s[5],s[6], s[7], s[8], s[9], s[10], s[11],s[12], s[13], s[14], s[15],
    s[16], s[17]], axis=0)
    return sss13

 ##Function for generating the SEIQRD based data the Pooled Bayesian model
```

```python
 def mymodel_B66p(R0_1,R0_2, T0_1,T0_2, T1_1,T1_2, T3_1,T3_2, e1_1,e1_2,
e2_1,e2_2, idx):
    N=14600000
    s1=[]
    allArrays=[]
    for i in range (int(idx)):
            if i<=9:
            RR=R0_1/T1_1
            bb=1/T0_1
            cc=1/T1_1
            dd=1/T3_1
            ee1=e1_1
            ee2=e2_1

            ss = model_SEIQRD(y0=yobs[15*i], theta=[RR, bb, cc, dd, ee1,
ee2])
            s1.append(ss)

        else:
            RR=R0_2/T1_2
            bb=1/T0_2
            cc=1/T1_2
            dd=1/T3_2
            ee1=e1_2
            ee2=e2_2
            ss = model_SEIQRD(y0=yobs[15*i], theta=[RR, bb, cc, dd, ee1,
ee2])
            s1.append(ss)
sss13 = theano.tensor.concatenate([s[0], s[1], s[2], s[3], s[4], s[5],s[6],
s[7], s[8], s[9], s[10], s[11],s[12], s[13], s[14], s[15], s[16], s[17]],
axis=0
return sss13

######   Preparing data for training
N=14600000                      # Population of Ontario
times=np.arange(1, 16, 1)     # Initiating time window of 15 days
Num_Train_Periods=18          # Number of training periods
x_train = Num_Train_Periods   # Number of training periods
y_train = yobs[0:270,:]       # Newly infected cases and mortality data

##The Hierarchical Bayesian model based on SEIQRD (HBN_SEIQRD)
with pm.Model() as model5hh:
    N=14600000                     # Population of Ontario
    n_bi_week1=10                  # biweekly periods in the 1st wave of the
outbreak
    n_bi_week2=8                   # biweekly periods in the 2nd wave of the
outbreak
    # Parameter's priors
      R0 = pm.Bound(pm.Normal, lower=1)("R0", 2.9, 2 )
      T0 = pm.Bound(pm.Lognormal,lower=1,upper=14)("T0", pm.math.log(5.5), 1)
```

```python
        T1 = pm.Bound(pm.Lognormal,lower=1,upper=14)("T1", pm.math.log(5.1), 1)
        T3 = pm.Bound(pm.Lognormal,lower=1,upper=30)("T3", pm.math.log(11.5),
1)
        e1 = pm.Bound(pm.Normal,lower=.5,upper=1)("e1", 0.8, .2)
        e2 = pm.Bound(pm.Normal,lower=0,upper=0.1)("e2", 0.01, .2)

        R0_1 = pm.Bound(pm.Normal, lower=1) ("R0_1", mu= R0, shape=n_bi_week1)
        R0_2 = pm.Bound(pm.Normal, lower=1) ("R0_2", mu= R0, shape=n_bi_week2)

        T0_1 = pm.Bound(pm.Lognormal,lower=1,upper=14) ("T0_1",
        mu=pm.math.log(T0), shape=n_bi_week1)
        T0_2 = pm.Bound(pm.Lognormal,lower=1,upper=14) ("T0_2",
        mu=pm.math.log(T0), shape=n_bi_week2)
        T1_1 = pm.Bound(pm.Lognormal,lower=1,upper=14) ("T1_1",
        mu=pm.math.log(T1), shape=n_bi_week1)
        T1_2 = pm.Bound(pm.Lognormal,lower=1,upper=14) ("T1_2",
        mu=pm.math.log(T1), shape=n_bi_week2)
        T3_1 = pm.Bound(pm.Lognormal,lower=1,upper=30) ("T3_1",
        mu=pm.math.log(T3), shape=n_bi_week1)
        T3_2 = pm.Bound(pm.Lognormal,lower=1,upper=30) ("T3_2",
        mu=pm.math.log(T3), shape=n_bi_week2)

        e1_1 = pm.Bound(pm.Normal, lower=0.5, upper=1)("e1_1", mu=e1,
shape=n_bi_week1)
        e1_2 = pm.Bound(pm.Normal, lower=0.5, upper=1)("e1_2", mu=e1,
shape=n_bi_week2)
        e2_1 = pm.Bound(pm.Normal, lower=0, upper=0.1)("e2_1",  mu=e2,
shape=n_bi_week1)
        e2_2 = pm.Bound(pm.Normal, lower=0, upper=0.1)("e2_2",  mu=e2,
shape=n_bi_week2)

        # Forecast using the SEIQRD model based on these parameters
        obs=mymodel_B66(R0_1,R0_2, T0_1,T0_2, T1_1,T1_2, T3_1,T3_2, e1_1,e1_2,
        e2_1,e2_2, x_train)
        Cases = pm.Lognormal("Cases", mu=pm.math.log(obs[:,[3,5]]),
        observed=y_train[:,[3,5]])

##The Pooled Bayesian model based on SEIQRD (Pooled_Bayesian_SEIQRD)
with pm.Model() as model5p:
    N=14600000
    n_bi_week1=10
    n_bi_week2=8

    # first Peak
    R01 = pm.Bound(pm.Normal, lower=1, upper=6)("R0", 2.9, 2, )
    T01 =  pm.Lognormal("T01", pm.math.log(5.5), 1)
    T11 =  pm.Lognormal("T11", pm.math.log(5.1), 1)
    T31 =  pm.Lognormal("T31", pm.math.log(11.5), 1)
    e11  = pm.Uniform("e11", 0.1, 1)
    e21  = pm.Uniform("e21", 0, 1)
```

```python
    # Second Peak
    R02 = pm.Bound(pm.Normal, lower=1, upper=6)("R02", 2.9, 2, )
    T02 =  pm.Lognormal("T02", pm.math.log(5.5), 1)
    T12 =  pm.Lognormal("T12", pm.math.log(5.1), 1)
    T32 =  pm.Lognormal("T32", pm.math.log(11.5), 1)
    e12  = pm.Uniform("e12", 0.1, 1)
    e22  = pm.Uniform("e22", 0, 1)

    obs=mymodel_B66p(R01,R02, T01,T02, T11,T12, T31,T32, e11,e12, e21,e22,
x_train)
    Cases = pm.Lognormal("Cases", mu=pm.math.log(obs[:,[3,5]]),
observed=y_train[:,[3,5]])
```

The UnPooled Bayesian model based on SEIQRD (UnPooled_Bayesian_SEIQRD)

```python
with pm.Model() as model5indv:
    N=14600000
    n_bi_week1=10
    n_bi_week2=8
    R0_1 = pm.Bound(pm.Normal, lower=.5, upper=6) ("R0_1", 2.9, 2,
shape=n_bi_week1)
    R0_2 = pm.Bound(pm.Normal, lower=.5, upper=6) ("R0_2", 2.9,
2,shape=n_bi_week2)

    T0_1 = pm.Lognormal ("T0_1", mu=pm.math.log(5.5),shape=n_bi_week1 )
    T0_2 = pm.Lognormal ("T0_2", mu=pm.math.log(5.5), shape=n_bi_week2)

    T1_1 = pm.Lognormal ("T1_1", mu=pm.math.log(5.1), shape=n_bi_week1)
    T1_2 = pm.Lognormal ("T1_2", mu=pm.math.log(5.1), shape=n_bi_week2)
    T3_1 = pm.Lognormal ("T3_1", mu=pm.math.log(11.5), shape=n_bi_week1)
    T3_2 = pm.Lognormal ("T3_2", mu=pm.math.log(11.5), shape=n_bi_week2)

    e1_1 = pm.Bound(pm.Normal, lower=0.1, upper=1)("e1_1", .80,
shape=n_bi_week1)
    e1_2 = pm.Bound(pm.Normal, lower=0.1, upper=1)("e1_2", 0.01,
shape=n_bi_week2)
    e2_1 = pm.Bound(pm.Normal, lower=0, upper=1)("e2_1",  .80,
shape=n_bi_week1)
    e2_2 = pm.Bound(pm.Normal, lower=0, upper=1)("e2_2",  .01,
shape=n_bi_week2)

    obs=mymodel_B66(R0_1,R0_2, T0_1,T0_2, T1_1,T1_2, T3_1,T3_2, e1_1,e1_2,
    e2_1,e2_2, x_train)
    Cases = pm.Lognormal("Cases", mu=pm.math.log(obs[:,[3,5]]),
    observed=y_train[:,[3,5]])
```

Codes for generating the visualization using graphviz

```python
    model5hh_graph=pm.model_to_graphviz(model5hh)
    model5hh_graph.view().save('image1.png')
    #model5indv_graph=pm.model_to_graphviz(model5indv)
```

```
      #model5indv_graph.view().save('image1.png')
      #model5p_graph=pm.model_to_graphviz(model5p)
      #model5p_graph.view().save('image1.png')


##Codes for generating Variational inference
# for Hierarchical method
with model5hh:
    inference = pm.ADVI()
    approxhh =
pm.fit(n=1000,method='advi',obj_optimizer=pm.adagrad(learning_rate=0.1))
# for unpooled model
with model5indv:
    inference = pm.ADVI()
    approxindv = pm.fit(n=1000,
method='advi',obj_optimizer=pm.adagrad(learning_rate=0.1))
# for the Pooled model
with model5p:
    inference = pm.ADVI()
    approxp = pm.fit(n=1000,
method='advi',obj_optimizer=pm.adagrad(learning_rate=0.1))

    #Ploting the inference (Fig 7.6: ELBO profile)
        plt.plot(approxhh.hist)
        plt.ylabel("Negative ELBO")
        plt.xlabel("iteration");

## MCMC Sampling using NUTS (by default in PYMC3)
with model5hh:
    approx_sample1=approxhh.sample(10000)
    approx_sample = approx_sample1[9500:]      # collecting most recent  500
samples
                                               assuming burning period is
over


## Model comparison for Table 7.4
df_comp_loo2 = az.compare({"model5hh": approx_samplehh,"model5indv":
approx_sampleindv, "model5p": approx_samplep})
df_comp_loo2


## Trace plots
    az.plot_trace(approx_sample)

## Extracting parameters of the HBN_SEIQRD model from the samples and
forecast
    #R0s=approx_sample[R0]
    R0_is1=approx_sample[R0_1]
    R0_is2=approx_sample[R0_2]
```

```
        T0s1=approx_sample[T0_1]
        T0s2=approx_sample[T0_2]
        T1s1=approx_sample[T1_1]
        T1s2=approx_sample[T1_2]
        T3s1=approx_sample[T3_1]
        T3s2=approx_sample[T3_2]
        e1s1=approx_sample[e1_1]
        e1s2=approx_sample[e1_2]
        e2s1=approx_sample[e2_1]
        e2s2=approx_sample[e2_2]

        ykc=[]    # Initialization of array for newly infected cases
        ykm=[]    # Initialization of array for mortality
        pp=6       # The most recent biweekly periods parameters, we have 7
biweekly periods
                    in the second wave [0, 1,2,3,4,5,6], so using pp=6

        for i in range(500):
            yk= odeint(SEIQRD, t=ttt, y0=yobs[15*pp],
            args=((R0_is1[i,ppp]/T1s1[i,ppp], 1/T0s1[i,ppp], 1/T1s1[i,ppp],
            1/T3s1[i,ppp], e1s1[i,ppp], e2s1[i,ppp]),),rtol=1e-6)
            ykc.append(yk[:,3])
            ykm.append(yk[:,5])

        ykcs=np.array(ykc)
        ykms=np.array(ykm)

        #Forecast of fatality
        ycal0_m_m=np.concatenate([yobs[0:270,5], np.quantile(ykms, .50,
axis=0)])
        ycal0_q75_m=np.concatenate([yobs[0:270,5], np.quantile(ykms, .75,
axis=0)])
        ycal0_q25_m=np.concatenate([yobs[0:270,5], np.quantile(ykms, .25,
axis=0)])

        #Forecast plots
        fig, ax=plt.subplots()
        t5=np.arange(0, 315, 1)  # Total period from 0 to the forecasted range

        ax.plot(t5,ycal0_m_m,color='k', label='most probable
value',linewidth=2.0)
        ax.plot(t5,ycal0_stdm_m,'k--', linewidth=2.0)
        ax.plot(t5,ycal0_stdp_m,'k--', linewidth=2.0)
        ax.scatter(t5, D1[0:315,1],s=8, color='r', label= 'true value')
        ax.fill_between(t5, ycal0_stdm,ycal0_stdp, where=t5>=270, alpha=0.25,
        label='standard deviation')
        sns.despine()
        ax.legend()
        ax.set(xlabel="Number of Days", ylabel="Cumulative Mortality");
```

```python
#Placing the model Forecast in the models reported by CDC
Death_Forecast1=pd.read_excel('Forecast_Arizona_Dec_Death_2.xlsx',
                    header= None, index_col=False)
            #Forecast_Arizona_Dec_Death_2 has the forecast value
            reported by distinct models in the studied period
Death_Forecast=Death_Forecast1.values

fig, ax=plt.subplots()
#ax.plot(CYC_m,fp_25_m, color='b', label='5 Percentile',linewidth=0.5)
ax.plot([t5[285],t5[314]],[ycal0[285],ycal0[314]],color='k',
label='HBN_SEIQRD',linewidth=2.0)
ax.plot([t5[285],t5[314]],[ycal0_stdm[285],ycal0_stdm[314]],'k--',
linewidth=2.0)
ax.plot([t5[285],t5[314]],[ycal0_stdp[285],ycal0_stdp[314]],'k--',
linewidth=2.0)
ax.fill_between([t5[285],t5[314]],
[ycal0_stdm[285],ycal0_stdm[314]],[ycal0_stdp[285],ycal0_stdp[314]],
alpha=0.30, label='std_HYC_BN_SEIQRD')

ax.plot([t5[285],t5[314]],[Death_Forecast[0,0],Death_Forecast[0,1]],col
or='b', label='BPagano',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[1,0],Death_Forecast[1,1]],col
or='g', label='UGA-CEID',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[2,0],Death_Forecast[2,1]],col
or='c', label='Columbia-UNC',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[3,0],Death_Forecast[3,1]],col
or='m', label='Covid19Sim',linewidth=2.0)

ax.plot([t5[285],t5[314]],[Death_Forecast[4,0],Death_Forecast[4,1]],'y'
, label='MIT-ORC',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[5,0],Death_Forecast[5,1]],'b-
-', label='Ensemble',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[6,0],Death_Forecast[6,1]],'g-
-', label='Columbia',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[7,0],Death_Forecast[7,1]],'c-
-', label='DDS',linewidth=2.0)

ax.plot([t5[285],t5[314]],[Death_Forecast[8,0],Death_Forecast[8,1]],'y'
, label='LSHTM',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[9,0],Death_Forecast[9,1]],'b-
-', label='Google-HSPH',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[10,0],Death_Forecast[10,1]],'
g--', label='GT-DeepCOVID',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[11,0],Death_Forecast[11,1]],'
c--', label='ISU',linewidth=2.0)

ax.plot([t5[285],t5[314]],[Death_Forecast[12,0],Death_Forecast[12,1]],'
y', label='JHU-IDD',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[13,0],Death_Forecast[13,1]],'
b--', label='JHU-APL',linewidth=2.0)
```

```python
ax.plot([t5[285],t5[314]],[Death_Forecast[14,0],Death_Forecast[14,1]],'
g--', label='Karlen',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[15,0],Death_Forecast[15,1]],'
c--', label='LANL',linewidth=2.0)

ax.plot([t5[285],t5[314]],[Death_Forecast[16,0],Death_Forecast[16,1]],'
y', label='Microsoft',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[17,0],Death_Forecast[17,1]],'
b--', label='MIT-LCP',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[18,0],Death_Forecast[18,1]],'
g--', label='MIT-CovAlliance',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[19,0],Death_Forecast[19,1]],'
c--', label='MOBS',linewidth=2.0)

ax.plot([t5[285],t5[314]],[Death_Forecast[20,0],Death_Forecast[20,1]],'
y', label='Oliver Wyman',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[21,0],Death_Forecast[21,1]],'
b--', label='PSI',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[22,0],Death_Forecast[22,1]],'
g--', label='RPI-UW',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[23,0],Death_Forecast[23,1]],'
c--', label='CovidComplete',linewidth=2.0)

ax.plot([t5[285],t5[314]],[Death_Forecast[24,0],Death_Forecast[24,1]],'
y', label='UA',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[25,0],Death_Forecast[25,1]],'
b--', label='UCLA',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[26,0],Death_Forecast[26,1]],'
g--', label='UCSB',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[27,0],Death_Forecast[27,1]],'
c--', label='UCSD-NEU',linewidth=2.0)

ax.plot([t5[285],t5[314]],[Death_Forecast[28,0],Death_Forecast[28,1]],'
y', label='UMass-MB',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[29,0],Death_Forecast[29,1]],'
b--', label='UM',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[30,0],Death_Forecast[30,1]],'
g--', label='USC',linewidth=2.0)
ax.plot([t5[285],t5[314]],[Death_Forecast[31,0],Death_Forecast[31,1]],'
c--', label='Wadhwani',linewidth=2.0)

sns.despine()
ax.legend(loc='center left',bbox_to_anchor=(1,0.5), ncol=4)
ax.set(xlabel="Number of Days", ylabel="Cumulative Mortality");
```

Appendix E: Pandemic forecast models displayed by CDC

https://www.cdc.gov/coronavirus/2019-ncov/science/forecasting/forecasts-cases.html
https://github.com/cdcepi/COVID-19-Forecasts/blob/master/COVID-19_Forecast_Model_Descriptions.md

Models			
BPagano	**Bob Pagano**	SIR model	Projections assume that the effects of interventions are reflected in the observed data and will continue going forward.
UGA-CEID	**University of Georgia, Center for the Ecology of Infectious Disease**	Statistical random walk model	Projections assume that social distancing policies in place at the date of calibration are extended for the future weeks.
Columbia-UNC	**Columbia University and University of North Carolina**	Statistical survival-convolutional model	This model assumes that transmission intensity will peak in early July and then gradually decline.
Covid19Sim	**Covid-19 Simulator Consortium**	SEIR model	Intervention assumptions: This model is based on assumptions about how levels of social distancing will change in the future. Hospitalization assumptions: The number of new hospitalizations per day are estimated from the number of infections, using state-specific hospitalization rates.
MIT-ORC		SEIR model	The projections assume that interventions will be reinstated if transmission reaches certain thresholds
Ensemble	**University of Massachusetts, Amherst**	The ensemble is a combination of 4 to 20 models, depending on the availability of forecasts for each location.	The ensemble forecasts include all submitted forecasts, derived from models that assume certain social distancing measures will continue and models that assume those measures will not continue.
Columbia	**Columbia University**	Metapopulation SEIR model	Intervention assumptions: This model assumes that contact rates will increase 5% during the first week of the forecast period. Following week 1, the reproductive number is then set to 1.0. Hospitalization assumptions: The model uses state-specific hospitalization data,

			when available. In states without hospitalization data, the model uses the national average value for hospitalization data.
DDS	**Discrete Dynamical Systems**	Bayesian hierarchical model	This model assumes that the effects of interventions are reflected in the observed data and will continue going forward.
LSHTM	**London School of Hygiene and Tropical Medicine**	This forecast is an ensemble of three different models: A time-varying reproductive number-base model, a time series model based on numbers of deaths, and a time series model based on numbers of cases and deaths.	Intervention assumptions: These projections assume that current interventions will not change during the forecasted period.
Google-HSPH	**Google and Harvard School of Public Health**	SEIR model fit with machine learning	These forecasts implement changes to future population mobility in order to predict COVID-19 transmission intensity.
GT-DeepCOVID	**Georgia Institute of Technology, College of Computing**	Deep learning	This model assumes that the effects of interventions are reflected in the observed data and will continue going forward.
ISU	**Iowa State University**	Nonparametric spatiotemporal model	These projections do not make any specific assumptions about which interventions have been implemented or will remain in place.
JHU-IDD	**Johns Hopkins University, Infectious Disease Dynamic Lab**	Metapopulation SEIR model	These projections assume that current interventions will not change during the forecasted period. Daily hospitalizations are estimated from predictions of daily cases. A standard proportion is applied to all states

JHU-APL	**Johns Hopkins University, Applied Physics Lab**	Metapopulation SEIR model	This model assumes that the effects of interventions are reflected in the observed data and will continue going forward.
Karlen	**Karlen Working Group**	Discrete time difference equations	Intervention assumptions: This model assumes that the effects of interventions are reflected in the observed data and will continue going forward. Hospitalization assumptions: The model uses state-specific hospitalization data. New hospitalizations are estimated from these data, or from the estimated number of new infections that will occur in each location.
LANL	**Los Alamos National Laboratory**	Statistical dynamical growth model accounting for population susceptibility	Intervention assumptions: This model assumes interventions in place on the first day of the forecast will remain in place for the next four weeks. Hospitalization Assumptions: State demographics and age-group symptomatic case hospitalization rates are used to estimate the daily number of hospitalizations, based on estimates of the total number of infections.
Microsoft	**Microsoft AI**	SEIR model on a spatiotemporal network	Intervention Assumptions: These projections assume that current interventions will not change during the forecasted period.
MIT-CovAlliance	**Massachusetts Institute of Technology, COVID-19 Policy Alliance**	SIR model	Intervention Assumptions: The projections assume that current interventions will remain in place indefinitely.
MOBS	**Northeastern University, Laboratory for the Modeling of Biological and Socio-technical Systems**	Metapopulation, age-structured SLIR model	Intervention assumptions: The projections assume that social distancing policies in place at the date of calibration are extended for the future weeks.
Oliver Wyman	**Oliver Wyman**	Time-dependent SIR model for detected and	Intervention assumptions: These projections assume that current

		undetected cases	interventions will not change during the forecasted period.
PSI	**Predictive Science Inc.**	Stochastic SEIRX model	**Intervention assumptions:** These projections assume that current interventions will not change during the forecasted period.
RPI-UW	**Rensselaer Polytechnic Institute and University of Washington**	SIR model fit to mobility data	**Intervention assumptions:** These projections calibrate the rate of transmission to the average rate of population mobility since the start of the epidemic and assume that this relationship will not change in the next four weeks.
CovidComplete	**Steve McConnell**	Statistical prediction model	This model assumes that the effects of interventions are reflected in the observed data and will continue going forward.
UA	**University of Arizona**	SIR model with data assimilation	This model assumes that current interventions will remain in effect for at least four weeks after the forecasts are made.
UCLA	**University of California, Los Angeles**	Modified SEIR model	Intervention assumptions: This model assumes that contact rates will increase as states reopen. The increase in contact rates is calculated for each state. **Hospitalization Assumptions:** The number of new hospitalizations per day are estimated from the number of infections, using state-specific hospitalization rates.
UCSB	**University of California, Santa Barbara**	An attention mechanism (deep learning) time series model	These projections assume that current interventions will not change during the forecasted period. The number of new hospitalizations per day are estimated from the number of infections, using state-specific hospitalization rates.
UCSD-NEU	**University of California, San Diego and Northeastern University**	Age-structured metapopulation model with deep learning	These projections assume that current interventions will not change during the forecasted period.

UM	**University of Michigan**	Ridge regression	These projections assume that current interventions will not change during the forecasted period.
USC	**University of Southern California**	SIR model	These projections assume that current interventions will not change during the forecasted period. **Hospitalization assumptions:** The number of new hospitalizations per day are estimated from the number of infections.
Wadhwani	**Wadhwani AI**	Bayesian SEIR model	These projections assume that current interventions will not change during the forecasted period.
IHME	**Institute of Health Metrics and Evaluation**	Combination of a mechanistic disease transmission model and a curve-fitting approach	Projections are adjusted to reflect differences in aggregate population mobility and community mitigation policies. Daily hospitalizations are estimated from predictions of daily deaths, using state hospitalization rates, where available.
Imperial	**Imperial College, London**	Ensembles of mechanistic transmission models, fit to different parameter assumptions	These projections do not make any specific assumptions about which interventions have been implemented or will remain in place.
CMU	**Carnegie Mellon University**	Autoregressive time-series model	These projections do not make specific assumptions about which interventions have been implemented or will remain in place.