Causal Inference and Interpretable Machine Learning for Multiscale Environmental Data Analysis

by

© Qiao Kang, M.Sc.

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 and Applied Science

Memorial University of Newfoundland

June 2024

St. John's Newfoundland and Labrador Canada

ABSTRACT

Environmental data analysis encompasses methods including domain specific environmental modelling, statistics, and data-driven methods (e.g., artificial intelligence) to interpret observational and experimental datasets for tackling environmental issues. The field of environmental data analysis has experienced significant advancement over the last decade, fueled by the exponential increase in data quantity and complexity and the progression of data-driven paradigms alongside artificial intelligence. This field faces several key challenges, including 1) the lack of means for analysis from causal perspectives, especially in complicated multivariable problems, 2) the relatively high computational cost associated with partial-differential-equation-based models that incorporate physics priors, compounded by intrinsic uncertainties in parameter tuning processes, and 3) the frequent situations with limited data available or valid for analysis. This dissertation research aims to bridge the gaps by developing a set of integrated methods that meld the strengths of interpretable machine learning and causal inference with classic tools for environmental data analysis and modelling. It entails the following major tasks: 1) to introduce an interpretable data analysis framework that leverages machine learning and causal inference. This framework can not only promote a deeper understanding of the causal relationships within environmental data but also serve as a testament to the value and potential of applying interpretative analytics in environmental fields. It is exhibited by a case study on the relationships between environmental factors and pandemic severity. 2) to develop a causal-prior embedded neural network, utilizing experimental data and parameters fitted from physics-based models, offering a systematic integration of lab experiments, physics-based simulation, causal inference techniques, and neural network

modelling. The method is demonstrated through an integrated experimental and modelling study on the fate and transport of metformin, an emerging contaminant, in a porous medium. 3) To propose and test a transfer learning-based method to estimate the occurrences of environmental pollutants released or closely associated with human activities under data-scarce scenarios, supported by a novel neural network architecture and a comprehensive model fine-tuning strategy. The method is exemplified through a global risk assessment of metformin with a special attention on Canadian ecozones and the Arctic and sub-Arctic regions to showcase the method's effectiveness in enhancing environmental risk evaluation in data-limited contexts.

The dissertation research advances the field of environmental data analysis by developing a set of new methodologies based on causal inference and interpretable machine learning. Those methods deliver benefits including enhanced model interpretability, reduced computational costs, and improved efficiency in dataset utilization, enabling robust analysis of environmental data across diverse scales. The research can offer not only robust and effect methodologies for actionable environmental data analysis and modelling but also enhance our capability to harness vast and complex environmental data for informed decision-making and policy development.

ACKNOWLEDGEMENT

I am deeply thankful for the privilege of pursuing my Ph.D. under the guidance and support of my supervisor, Dr. Bing Chen. As my mentor, Dr. Chen has been an unwavering source of support, offering invaluable guidance, freedom to explore, and extensive knowledge in the interdisciplinary field of environmental engineering and science, which significantly boosted my ability and motivation to navigate through complex research challenges. I am also sincerely grateful to my co-supervisors, Dr. Baiyu (Helen) Zhang, and Dr. Yuanzhu (Peter) Chen, for their invaluable advice and steadfast support. Dr. Zhang's deep commitment and heartfelt support towards me and my peers have been instrumental in my academic and personal growth. Her genuine care and readiness to help have made a profound impact. Dr. Yuanzhu (Peter) Chen has provided me with countless invaluable insights and opportunities to refine my expertise in machine learning and computer science, significantly broadening my horizons.

I gratefully acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Foundation for Innovation (CFI) for their generous funding of our research projects. Additionally, I am grateful to the Digital Research Alliance of Canada for providing cloud computing resources. I also deeply appreciate the opportunity provided by Memorial University, where I have spent the most fulfilling and joyful time of my life.

I extend my heartfelt appreciation to Lidan Tao, whose assistance and dedication have been instrumental in advancing my research. I would also like to extend my special thanks to my colleagues: Dr. Xudong Ye, Dr. Xing Song, Ethan Matchinski, Dr. Bo Liu, Dr. Yiqi

Cao, Dr. Min Yang, Dr. Xixi Li, Yifu Chen, Dr. Zhiwen (Joy) Zhu, Dr. Arpana A. Datta, Jingjing Ling, Dr. Weiyun Lin, Dr. Qianqian Dong, Dr. Hongjing (Derek) Wu, Dr. Liang (Liam) Jing, Yunwen Tao, Dr. Xiujuan Chen, Dr. Fuqiang Fan, Dr. Wendy Huang, Dr. Qinhong (Tammy) Cai, and many others for their friendship, support, and camaraderie. I consider myself incredibly fortunate to be a part of such an inspiring community. To everyone who has supported me throughout my Ph.D. program, whom I have not had the chance to name individually, I express my sincerest gratitude. Your support has deeply lightened this journey, transforming it into a profound experience.

Lastly, I reserve my deepest gratitude for my family, particularly my mother. Your constant support, encouragement, and faith in my abilities have fortified my strength and resilience, making this PhD journey possible.

CO-AUTHORSHIP STATEMENT

I, Qiao Kang, hold the principal author status for all the manuscript chapters (Chapters 2-5) in this dissertation. Each chapter is a collaborative effort with my supervisors and co-researchers whose contributions have been instrumental in shaping this work. The specifics of these collective efforts are outlined below.

In Chapter 2, content on environmental modelling methods is derived from work co-authored with Datta, A. R., Chen, B., and Ye, X., available at [https://doi.org/10.1016/bs.amb.2018.09.002]. I conceptualized the study with Datta, A. R., conducted literature reviews, and participated in writing and editing. I prepared the initial draft of the manuscript, which was then refined through collaborative revisions.

Chapter 3 is derived from a study published in collaboration with Song, X., Xin, X., Chen, B., Chen, Y., Ye, X., and Zhang, B., available at [https://doi.org/10.1021/acs.est.1c02204]. I conceived the study with the input from Song, X., Xin, X., Ye, X., Zhang, B. and Chen, B. I, Song, X., and Xin, X. contributed to data collection and processing. I, together with Chen, Y., developed the program script and conducted the data analysis. I wrote the first version of the manuscript. Zhang, B., Chen, Y. and Chen, B. contributed to subsequent revisions.

Chapter 4 is based on collaborative research with Chen, B., Cao, Y., Song, X., Ye, X., Li, X., Wu, H., and Zhang, B., titled "Causal Prior-Embedded Physics-Informed Neural Networks and a Case Study on Metformin Transport in Porous Media," Accepted by *Water Research*. I conceived the study with the input from Song, X., Zhang, B. and Chen, B. I, Cao, Y., and Song, X. contributed to conducting the transport experiments and gathering

data. I developed the program scripts for processing, analyzing, and visualizing the data along with Ye, X., and Wu, H. Visualization and interpretation of the results were further enhanced by the contributions from Cao, Y., Song, X., and Li, X. The initial manuscript draft was written by me, with subsequent feedback provided by Zhang, B., Chen, Y. and Chen, B.

Chapter 5 builds upon and extends the work presented in "*Mapping the Global Environmental Risk of Metformin: A Transfer Learning Approach*," a manuscript co-authored with Yang, M., Song, X., Cao, Y., Liu, B., Ye, X., Wu, H., Zhang, B. and Chen, B. I conceived the study with the input from Zhang, B. and Chen, B. I designed the transfer learning-based framework and the EffluentNet structure. I developed the program scripts for processing, analyzing and visualizing the data along with Ye, X., and Wu, H. Visualization and interpretation of the results were further enhanced by the contributions from Cao, Y. and Song, X. The manuscript's refinement benefited from the comments and discussions from Chen, Y., Zhang, B., and Chen, B.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACTV	Daily Degree of Activeness
AI	Artificial Intelligence
AQI	Air Quality Index
ATE	Average Treatment Effect
BED	Hospital Beds per Thousand Population
DAG	Direct Acyclic Graph
DDD	Defined Daily Dose
DOC	Registered Medical Doctors per Thousand Population
FNFNES	First Nations Food, Nutrition & Environment Study
GDP	Gross Domestic Product
GNN	Graphical Neural Network
GPU	Graphics Processing Unit
HMD	Relative Humidity
IDF	International Diabetes Federation
NRS	Registered Nurses per Thousand Population
OECD	Organisation for Economic Co-operation and Development
PCA	Principal Component Analysis
PINN	Physics-Informed Neural Network
PLB	Placebo Refuter
POP	Population
PPCP	Pharmaceutical and Personal Care Products
PRES	Atmospheric Pressure

PRIM	Primary Sector of Gross Domestic Product
RCC	Random Common Cause Refuter
RQ	Risk Quotient
SAM	Structural Agnostic Model
SCM	Structural Causal Model
SEC	Secondary Sector of Gross Domestic Product
TEMP	Average Air Temperature
TERT	Tertiary Sector of Gross Domestic Product
TVLR	Inbound Travellers from Wuhan
UOC	Unobserved Confounder
WSPD	Wind Speed
WWTP	Wastewater Treatment Plant

CHAPTER 1

INTRODUCTION

1.1 Background

Environmental data analysis is a field dedicated to examining observational and experimental datasets to extract insights critical for addressing pressing environmental challenges. This discipline utilizes specialized modelling tools informed by prior knowledge, alongside statistical methods, to guide data interpretation and inform solutions (Gibert et al., 2018). Fueled by the exponential increase in environmental data volume and complexity and the progression of data-driven methods such as artificial intelligence, the field of environmental data analysis has experienced significant advancement over the last decade (Fleming et al., 2021; Rolnick et al., 2023). Leveraging advancements in computational power and data-driven techniques, the field of environmental data analysis possesses unique characteristics. A key aspect is the emphasis on understanding the reasons behind environmental phenomena, which is often as crucial, if not more so, than the phenomena themselves (Carriger et al., 2016). This is because environmental issues typically span multiple disciplines, such as chemistry, biology, geoscience, and epidemiology, and so on. This interdisciplinary nature can make the underlying mechanisms of observations complex and not as straightforward as those seen in controlled laboratory settings, raising the risk of generating misleading conclusions solely based on correlations (J. Zhu et al., 2023). Moreover, the inherent uncertainty and complexity of environmental issues require engineers and scientists to not only rely on traditional physics-based models but also to embrace data-driven approaches (Beven, 2007). These approaches are essential for extracting insights from a wide range of data sources and formats, both experimental and observational. For example, data-driven research has yielded significant findings in environmental research field, from estimating arsenic levels

in global groundwater to identifying significant evidence of climate change in a large amount of studies (Podgorski & Berg, 2020; Callaghan et al., 2021), highlighting the value of further exploration in this area. Furthermore, some urgent environmental issues are emerging, presenting critical challenges with limited data availability, such as the case with emerging pollutants (Arpin-Pont et al., 2016; Archer et al., 2017). These substances are not routinely monitored worldwide, leading to the datasets that are sparse and largely confined to limited numbers of studies without systematic collection efforts, comparing with ordinary pollutants. Therefore, we are facing significant challenges and environmental data analysis can play a critical role in addressing those problems amid changing conditions, underscoring the urgent need for innovative and effective analytical strategies to utilize the limited available data.

In modern environmental data analysis, two primary modelling paradigms are distinguished: physics-based models and data-driven methods (Šimůnek & van Genuchten, 2008; Zhong et al., 2021). From a methodological perspective, physics-based models have been a valuable tool and extensively applied in the environmental field. Although these models are reliable and useful, they are characterized by high computational demands and a complex recalibration process which reduce its applicability. This limits their application for rapid analyses, especially when computational resources are scarce and time is of the essence, such as in emergency decision-making processes (Šimůnek & van Genuchten, 2008; Ye et al., 2020). Conversely, with the recognized power and effectiveness, the data-driven methods have been given growing attention along with concerns on their lack of transparency in many algorithms and limitations in reflecting the physical mechanisms or processes underlying the data (Zhong et al., 2021; J. Zhu et al., 2023). Therefore, given the

unique demands of the field and the inherent challenges within commonly used methodologies, some novel approaches beyond standard machine learning techniques and physics-based modelling methods are worth exploring.

Causal inference aims to investigate the underlying reasons behind data correlations, evaluating essential causal relationships for informed decision-making and policy development (Pearl, 2000). It enables researchers to investigate the mechanisms or driving factors behind specific phenomena, such as the association between air pollution and public health (Davis et al., 2022; Forster et al., 2020; G. He et al., 2020). Additionally, there is a growing demand for interpretable data-driven methods such as Physics-Informed Neural Networks (Zhong et al., 2021), which incorporates physics-based models with machine learning. This integration ensures that models are grounded in scientific principles, improving their interpretability and reliability while reducing their dependency on extensive datasets (Bandai & Ghezzehei, 2022; Cai et al., 2021). Furthermore, the adoption of techniques such as transfer learning in applied science and engineering illustrates its effectiveness in utilizing existing knowledge and data to tackle new, often data-limited, challenges (S. J. Pan & Yang, 2010). By adapting models trained for one task to another related one, transfer learning circumvents the need for large datasets traditionally required for training models from the ground up. Its utility in environmental engineering is particularly valuable, given the field's frequent encounters with sparse or difficult-to-acquire datasets (Cao et al., 2022; Chen et al., 2021). This approach not only speeds up research efforts but also improves prediction accuracy and model robustness, facilitating more effective environmental management and policymaking. It holds promise for addressing emergent yet data-sparse issues, like the presence and ecological risk of

pharmaceuticals and personal care products (PPCPs) in aquatic environments, a problem which has been widely recognized for its significance yet hampered by insufficient data (Wilkinson et al., 2022). Therefore, the aforementioned methods warrant further exploration to enhance and refine current methodologies in environmental data analysis, enabling more effective tackling of urgent environmental challenges.

1.2 Statement of Problems

The integration of machine learning into environmental data analysis heralds a new era of vast potential, promising to enhance our understanding and management of environmental systems. However, a notable challenge lies in environmental data analysis, specifically the gap in interpretative, data-driven methodologies and optimal data operational practice. This deficit profoundly affects our ability to fully harness the power of AI and other data-driven methods in addressing environmental concerns, specifically in four critical areas: causal interpretation of observational data, adaptation of physics-based models, pervasive issues of data scarcity, and optimal data curation practices.

(1) Need for causal inference framework for environmental data analysis

The task of extracting causal insights from observational data presents a significant challenge in environmental data analysis. While some problems may appear straightforward initially, a comprehensive investigation is often required to adequately assess causal links. The current analytical frameworks' inability to offer causal interpretations underscores a pressing need for innovative methods that can transform observational data into causal insights. The COVID-19 pandemic exemplifies this, as researchers have endeavored to determine the impact of potential environmental factors, such as meteorological conditions and air pollution, on the severity of the disease (Bashir,

Ma, Bilal, Komal, Bashir, Tan, et al., 2020). However, only a handful of studies have approached these associations from a causal perspective (Mastakouri & Schölkopf, 2020).

(2) Inherent challenges in physics-based models

Although models based on physics prior-knowledge are reliable and useful, they are marked by their intensive computational demands and the inherent uncertainty, especially from the parameterization processes, which significantly increase computation time and diminish their applicability (Raissi et al., 2019). The limitation renders these models less suitable for quick analyses in situations when computational resources are limited, and swift decision-making is vital such as during environmental emergency responses (Šimůnek & van Genuchten, 2008; Ye et al., 2020). Furthermore, these models encounter obstacles including equifinality, a phenomenon where diverse parameter sets produce indistinguishable output curves, raising the risk of misinterpretation (Beven, 1996). This issue was prominently illustrated in the studies investigating the fate and transport behaviours of emerging pollutants within porous media, where numerous parameter combinations derived from extensive parameterization processes yielded identical, well-fitted breakthrough curves, yet insufficient prior knowledge failed to discern among these combinations (Bandai & Ghezzehei, 2021). Stiffness is also a situation in which certain parameters lead to equations that pose significant challenges in solving (Asaro & Lubarda, 2006; Um et al., 2019). Lastly, a significant challenge in utilizing physics-based models in environmental engineering is the often-ambiguous interpretability of some parameters. These parameters are typically inferred by fitting models to observed data, leading to significant research efforts focused on curve-fitting (Kumar, 2012).

(3) Scarce data for emerging environmental challenges

Data scarcity is a pervasive challenge across various domains of applied science. The lack of comprehensive, high-quality datasets can significantly constrain the ability of researchers to draw meaningful conclusions and formulate effective management strategies (Gibert et al., 2018). This limitation becomes particularly apparent when investigating emerging pollutants, such as PPCP. Their widespread presence and ecological impacts are increasingly alarming. However, most of them are not covered by regulated monitoring programs, resulting in a lack of comprehensive data from both spatial and temporal scales. Efforts to understand the environmental footprint of PPCPs often need to grapple with the dual challenges of data insufficiency and the absence of methodologies designed to extract maximum insights from limited datasets (Wilkinson et al., 2022).

1.3 Objectives and Tasks

This dissertation research aims to develop an integrated framework aided by a set of interpretable machine learning and causal inference methods for effective data analysis at multiple scales, to provide more solutions for environmental studies. Specifically, it entails the following research tasks:

(1) To develop an interpretable data analysis framework that utilizes interpretable machine learning techniques and causal inference methodologies in environmental engineering and science, with an emphasis on promoting model interpretability and data transparency. This framework is to be exemplified by an in-depth analysis of the impact of environmental factors on COVID-19 severity, to illustrate the effectiveness and applicability of causal inference in environmental data analysis.

(2) To develop a causal-prior embedded neural network based on experimental data and physics-model fitted parameters to enhance the current environmental data analysis methodologies by systematically integrating experimental data, physics-based modelling, causal inference techniques, and neural networks. This method will be tested by an investigation on the fate and transport of metformin, a representative PPCP, in porous media.

(3) To develop a transfer learning-based methodology designed to estimate environmental pollutant occurrences with limited data, particularly emerging pollutants closely linked to human activities. This approach will be supported by a suite of complementary data science techniques, including an innovative neural network architecture and a comprehensive model fine-tuning strategy. The method is to be demonstrated through the global risk assessment of metformin, with a particular emphasis on Canadian ecozones and the Arctic and sub-Arctic regions, for its effectiveness on tackling the issue of data scarcity in environmental risk analysis.

1.4 Structure of the Dissertation

This dissertation is structured in a manuscript-based format, systematically unfolding through a series of research work and outcomes, to address the critical challenges in environmental data analysis. Our methods have been exemplified through case studies that illuminate the intricate interactions between environmental factors and the COVID-19 pandemic (Chapter 3), explore the behavior and movement of the emerging pollutant metformin (Chapter 4), and evaluate its global ecological risks (Chapter 5). A schematic overview of the research framework is depicted in Figure 1.1, illustrating the dissertation's structure and logical connections is organized as follows:

Chapter 2 provides a comprehensive review of key topics, laying the groundwork with a focus on emerging environmental health challenges. It introduces the methods employed in the research, including machine learning, causal inference, PINN (Physics-Informed Neural Network), and transfer learning, establishing a foundation for the subsequent analyses. Additionally, it provides a background review that details the intricate relationship between COVID-19 and air pollution and surveys the prevalence of metformin as an emerging pollutant in global waterbodies.

Chapter 3 introduces a causal inference framework utilizing the Structural Causal Model (SCM) and machine learning techniques and showcases its capability by investigating the potential causal relationships between COVID-19 severity and environmental factors across 166 Chinese cities, categorized into three clusters based on socio-economic characteristics. The chapter concludes with a comprehensive robustness check to examine the reliability of the potential causal links.

Chapter 4 proposes a novel transport modelling approach that incorporates experimentally derived causal priors into neural networks, with the aid of causal inference techniques. Its capability is demonstrated by a case study exploring the transport dynamics of metformin in sandy media. The transport characteristics of metformin and the effectiveness of the methodology are summarized and discussed.

Chapter 5 develops a modelling strategy ideal for scenarios where limited, yet valuable prior knowledge is available, utilizing transfer learning and semi-synthetic datasets. It also introduces EffluentNet, a novel neural network architecture specifically designed to simultaneously predict two or more associated distributions, demonstrating its applicability in scenarios where the relationships between distributions are known to be interconnected.

Its capability is demonstrated through a case study investigating the global distribution of metformin, identifies critical areas of concern and highlights the impact of pharmaceutical pollutants in pristine environments like Canadian ecozones and the Arctic and sub-Arctic regions. This chapter advocates for the adoption of culturally sensitive policies, particularly in regions inhabited by indigenous communities, to ensure environmental preservation that aligns with the socio-cultural fabric of those communities.

Chapter 6 concludes the dissertation by summarizing the key contributions and findings of the research. It underscores significant insights and makes recommendations for future research in environmental data analysis.

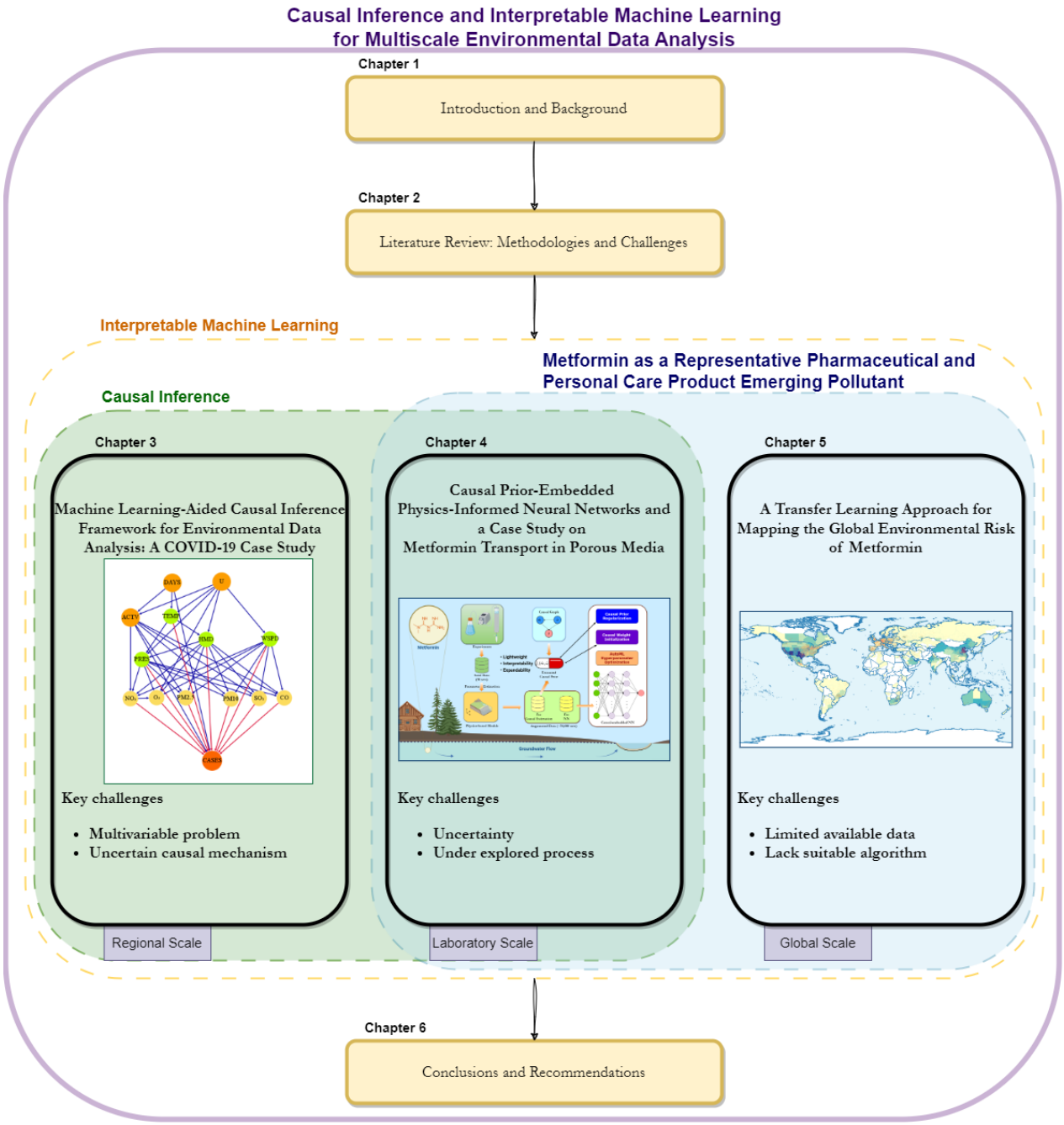


Figure 1.1 A schematic diagram of the dissertation research.

CHAPTER 2

LITERATURE REVIEW

2.1 Environmental Data Analysis[1]

Environmental data analysis encompasses techniques such as classic environmental modelling, statistics and artificial intelligence to analyze and interpret complex datasets, aiming to support environmental mitigation and decision-making (Gibert et al., 2018). As a complex and evolving field, environmental data analysis has seen rapid progress due to the substantial growth in both the volume and complexity of data generated by advancements in environmental analytical tools, monitoring technologies, and the availability of open datasets (Zhong et al., 2021). Consequently, the sources of data for environmental analysis involve not only traditional environmental variables, such as monitoring and observational data, satellite imagery, but also extend to unconventional data pertinent to environmental challenges (Wilkinson et al., 2022; Podgorski & Berg, 2020), such as public health indices and financial statements that highlight a company's carbon footprint (Callaghan et al., 2021; Heberling et al., 2021; Rolnick et al., 2023). Therefore, in a broad sense, environmental data analysis is characterized by its goals rather than the specific types of data it utilizes. This section reviews the current state of environmental data analysis and common challenges it can solve, aiming to provide a thorough background on the subject.

Before the era of 'big data', researchers primarily utilized two approaches to simulate and investigate environmental processes: 1) empirical models, exemplified by the Universal

[1] This section is partially based on and expanded from the following paper:
Datta, A. R., **Kang, Q.**, Chen, B., & Ye, X. (2018). Fate and transport modelling of emerging pollutants from watersheds to oceans: a review. Advances in marine biology, 81, 97-128.
Role: I conceptualized the study with Datta, A. R., conducted literature reviews, and participated in writing and editing. I prepared the initial draft of the manuscript, which was then refined through collaborative revisions.

Soil Loss Equation (Hudson, 1993) and 2) analytical and theoretical methods, such as the Navier–Stokes equations (Bear, 2013). Many of these models remain relevant today. Taking Hydrus-1D as an example, it is one of the modelling tools capable of simulating the transport of chemicals. It is a well-established software tool that incorporates a variety of physics-based equations, such as Richards' equation for saturated-unsaturated water flow and Fickian-based advection-dispersion equations for solute transport, enabling analysis of water flow and solute transport in variably saturated porous media. It has been used to investigate the fate and transport of substances such as propranolol, ciprofloxacin, clomipramine, caffeine and carbamazepine (Feizi et al., 2021; Koroša et al., 2020). The governing flow and solute transport equations were inversely solved via the finite element method from the observed breakthrough curves. This approach allows for the estimation of parameters related to porous media hydraulics and solute transport. However, like many other partial differential equation (PDE) models, it faces challenges such as equifinality, where different parameter combinations yield identical fitted curves, potentially leading to inaccurate interpretations (Beven, 1996), and stiffness, where reasonable parameters result in equations that are difficult to solve (Asaro & Lubarda, 2006). Thus, there is room for these models to benefit from advancements in modern data-driven methods.

The scenarios that environmental data analysis faces are also becoming increasingly complex. One example is the risk posed by emerging pollutants, which include PPCPs, deodorizers, fragrances, flame retardants, industrial chemicals, natural hormones and steroids, current-use pesticides, plasticizers, and surfactants (Diamond et al., 2011). The chemicals are not commonly monitored under existing environmental regulations but considered to pose potential risks to ecosystems and humans (Geissen et al., 2015). Due to

the rapid development of technology, industry, and resources, the increase in the number of emerging pollutants is alarming in the environment from the atmosphere to the subsurface and oceans. More than 1,000 emerging pollutants in the European aquatic environment have already been listed by the Norman Network, Network of reference laboratories, research centres and related organisations for monitoring emerging environmental substances (2016). How the CECs enter the environment depends on their usage patterns and application modes (La Farré et al., 2008). Emerging pollutants from urban and industrial sources are discharged into sewers and wastewater treatment plants. The removal of these pollutants is challenging since conventional treatment processes are usually not capable of eliminating them. For example, the removal of endocrine-disrupting compounds (EDCs) by lime softening or by coagulation by alum/ferric sulphate can be < 20% (Deblonde et al., 2011). Hence, those wastewater effluents can be considered as point source pollution to the water bodies. Once they reach the water bodies, they can be further transported downstream in dissolution form or present in the suspended solids form. The physicochemical properties of emerging pollutants, such as solubility, vapour pressure and polarity, determine their environmental behaviour (La Farré et al., 2008). Depending on the transport properties, some potential fate and transport pathways for emerging pollutants include leaching, surface runoff, and sorption (Geissen et al., 2015). Biological and chemical degradation may occur during the transportation to ambient water bodies as well. Eventually, the concentration of emerging pollutants in water bodies may vary from a low ng/L level to a mid μg/L level (Ahmed et al., 2017). Those variations in their occurrences are due to different doses applied in the treatment processes in various regions and also the inconsistent treatment efficiency of the wastewater treatment plants. Thus, the

stability of the physics-based models when modelling transport processes with significant concentration variations cannot be guaranteed. On the other hand, the relatively unknown characteristics of emerging pollutants exacerbate the issue of equifinality since such uncertainty makes it difficult to select the most reasonable parameter combinations from a series of candidates due to a lack of comprehensive knowledge about these pollutants. Lastly, the emerging nature of these pollutants often means that there is limited data available for robust process simulation and predictive modelling. The scarcity of data poses a significant challenge in assessing the environmental impact of emerging pollutants. The challenges outlined above underscore the potential for physics-based models to greatly benefit from the integration of modern data-driven methodologies.

Another prevalent scenario in environmental data analysis involves the exploration of causal relationships, as illustrated by the case between environmental factors and COVID-19. The global health crisis instigated by the COVID-19 pandemic has necessitated a critical examination of various contributing factors to its spread and severity. Among these, the correlation between environmental factors—specifically air pollution and meteorological conditions—and COVID-19 outcomes has garnered significant scholarly attention (F. Liu et al., 2021; X. Zhang et al., 2021). Some studies have highlighted the inadvertent environmental benefits arising from lockdown measures implemented worldwide (Adams, 2020; G. He et al., 2020; Lovrić et al., 2020). A pivotal study conducted across 325 cities in China revealed a substantial improvement in air quality, quantified by a 12.2% reduction in the Air Quality Index (AQI). This improvement varied across different pollutants and was more pronounced in northern cities, areas with higher income, and more industrialized regions (M. Wang et al., 2020). The relationship between

short-term exposure to air pollution and COVID-19 infection rates has been explored in various studies. One such study in China found positive associations between several air pollutants (PM2.5, PM10, NO_2, and O_3) and the incidence of COVID-19 cases. Interestingly, SO_2 exhibited a negative association with confirmed cases (Y. Wang et al., 2020). Similarly, many studies reported reductions in air pollution during the pandemic and interpreted such drop as a result of the lockdown policies implemented worldwide. Such interpretations are considered to be credible and reasonable (Adams, 2020; Cole et al., 2020; F. Liu et al., 2021; J. Liu et al., 2020; Lovrić et al., 2020; J. Wang et al., 2021; M. Wang et al., 2020; Y. Wang et al., 2020). Meanwhile, those reported links lead to the speculation that some causal mechanisms may exist behind the associations with several plausible hypotheses proposed. For instance, low wind speeds may facilitate the suspension of infectious particles in the air, potentially increasing exposure risks (Coccia, 2020a, 2021b). Additionally, exposure to air pollution may weaken individuals' immune systems, thereby elevating the infection rate. Exposure to air pollution may compromise people's immune systems and induce a higher infection rate (Kutter et al., 2021; Srivastava, 2021; Tian et al., 2021). Though no consensus has been reached, researchers deployed various approaches on several types of observational data and discussed the possibilities of some causal links' existence (Islam et al., 2021; Qu et al., 2020; Sunyer et al., 2021), even though some studies within the field have interpreted data with less caution, resulting in some confusion. Such a scenario highlights the complexity of environmental challenges caused by emerging pathogens and the multifaceted nature of the data they produce, underscoring the indispensable need for causal inference in environmental data analysis.

2.2 Causal Inference with Structural Causal Model

The Structural Causal Model (SCM) evolves from the Bayesian Network and Structural Equation Model. The main improvement of SCM compared with its predecessors is that SCM uses a causal diagram as part of the input. In this way, the prior knowledge is introduced into the system in a causal directional manner rather than bidirectional probability distribution in Bayesian Network and Structural Equation Model. Hence, the relationships in SCM can more accurately represent real-world causal links.

SCM uses a directed acyclic graph (DAG) to reflect the causal relationship between different variables. A variable in the dataset is a vertex in the graph, and a directed edge (arc) indicates a causal link. This causal diagram explicitly introduces prior knowledge regarding the data-generating process to the system. Given a causal diagram of a problem based on a series of mathematically proven graphic-based operations, a set of variables in the graph can be picked from all the given variables while following the graph-based operations. The selected variables are then sufficient to calculate the causal effects of interest. The set of these variables is hence called the sufficient set. Another important assumption is that if a causal effect can be estimated, all the variables in the sufficient set should be observable.

To have a deeper insight into causal inference, two conditional distributions that one might want to estimate during data analysis should be distinguished. The two distributions are given below (Pearl, 2000, 2014):

Observation p(y|x): The conditional distribution of Y when the variable X has the value x.

Intervene p(y|do(x)): The conditional distribution of Y when the variable X is set to x.

Though being similar, the two distributions are totally different. Only the second one can answer causal problems. Intervention means changing the value of a causal variable X on purpose and then observing the changes in the corresponding variable Y. The effect of an intervention operation expressed as a probability distribution is given as $P\big(y\big|do(x)\big)$. Mathematically, the ultimate goal of SCM is to estimate this distribution, which is hardly feasible in most real-world cases.

The primary goal of do-calculus is to estimate $P\big(y\big|do(x)\big)$ based on observed data outside of a controlled randomized experiment. For better comprehension, below is an axiomatic system for converting probability formulas containing the do operator with ordinary conditional probabilities. Let G be the directed acyclic graph associated with a causal model and let $P(\cdot)$ denote the probability distribution induced by the model. $G_{\overline{X}}$ indicates a modified graph G with all the edges pointing towards X are removed. Similarly, $G_{\underline{X}}$ indicates that all the outgoing edges from X are removed. For any disjoint subsets of X,Y,Z, and W, the following rules apply (Y. Huang & Valtorta, 2012):

Rule 1: Insertion/deletion of observations

$$P\ (y|\hat{x},z,w) = \ P\ (y|\hat{x},w)\ if\ (Y\ \perp\ Z|X,W\ )_{G_{\overline{X}}} \tag{2.1}$$

Rule 2: Action/observation exchange

$$P\ (y|\hat{x},\hat{z},w) = \ P\ (y|\hat{x},z,w)\ if\ (Y\ \perp\ Z|X,W\ )_{G_{\overline{X}\underline{Z}}} \tag{2.2}$$

Rule 3: Insertion/deletion of actions

$$P\ (y|\hat{x},\hat{z},w) = \ P\ (y|\hat{x},w)\ if\ (Y\ \perp\ Z|X,W\ )_{G_{\overline{X},\overline{Z(W)}}} \tag{2.3}$$

where Z(W) is the set of Z-nodes that are not ancestors of any W-node in $G_{\overline{X}}$. All the do-calculus-based operations are based on the three rules mentioned above.

Where Y is the outcome, T is the treatment, and W is a set of identified backdoor variables.

Given a sufficient set S, the causal effect can be expressed as:

$$P\big(Y\big|\, do(X = x)\big) = \sum_{s} P(Y = y|\, X = x, S = s)P(S = s) \tag{2.4}$$

Where $do(\cdot)$ indicates the intervention operation, X, Y indicate the treatment and the outcome variables, respectively. S indicates the variables in the sufficient set. In contrast, x, y, and s indicate the individual values in corresponding variables.

Another presentation of causal effect can be given in the form of average treatment effect (ATE), as described as follows:

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)] \tag{2.5}$$

Where $Y(1)$ represents the value of an outcome transport variable when the causal transport parameter is altered (i.e., the treatment is applied), and $Y(0)$ denotes the value of the outcome variable during the substance transport if the causal transport parameter is not changed (i.e., the treatment is not applied). $\mathbb{E}[\,\cdot\,]$ denotes the expectation over all simulated transport processes (Rubin, 1974).

Refutation techniques are commonly used to test the robustness of the causal estimators during experiments, which can be helpful testing the robustness of the assumed causal links

(Sharma & Kiciman, 2020). Refutation methods are a series of statistical experiments to assess the constructed DAG and the estimators' ability to withstand scrutiny, i.e., robustness. Some of these techniques include 1) the placebo treatment refuter, which involves randomizing the treatment variable affecting the outcome and expects the causal estimates to go to zero; 2) the adding random common cause refuter, which introduces an independent variable to test the sensitivity of causal estimates to a new common cause, with robust estimates remaining unchanged; and 3) the adding unobserved common causes refuter, which investigates the effect of adding a treatment-correlated confounder, indicating the presence of hidden confounders if the estimates change significantly (Sharma & Kiciman, 2020). For instance, the final effect estimates for downstream analysis can be chosen based on the most robust estimator between machine learning and linear estimations, factoring in the results of refutations. If both estimations yielded similar results, Occam's Razor principle was used, and hence the study opted for linear estimates. The extracted causal insights from the dataset can be further encapsulated into a series of causal functions.

One limitation of SCM is its reliance on the correctness of the graph model, which means it should correctly include causal relevant variables and potential confounders. On the other hand, challenges can also arise from the complexity of depicting intricate causal relationships in large multivariate systems with potential non-linear causal effects. These constraints necessitate cautious application and interpretation of SCM findings, especially when generalizing across different contexts (Pearl, 2000).

2.3 Machine Learning: Interpretable Models, Physics-Informed Neural Networks, and Transfer Learning

Machine learning methods suitable for analyzing environmental datasets should 1) excel in solving complex high-volume continuous data regression problems, 2) be capable of tackling the over-fitting problem, and 3) have a highly interpretable structure for research purposes. Due to these three demands, tree-based algorithms, including well-known models such as Random Forest, LightGBM, and XGBoost seemed to be more suitable for this task compared to other algorithms in the field (T. Chen & Guestrin, 2016; Ke et al., 2017).

Random Forest is a widely recognized and powerful ensemble learning technique used extensively in the field of machine learning for both classification and regression tasks (Breiman, 2001). As an ensemble of decision trees, Random Forest combines the predictions of multiple tree models to improve accuracy and reduce the risk of overfitting. This is achieved by training each tree on a random subset of the dataset and averaging their predictions for regression tasks or using a majority vote for classification. For environmental scientists and engineers, Random Forest serves as an indispensable tool for diverse applications, including but not limited to predicting PM 2.5 concentration (Zamani Joharestani et al., 2019) and evaluating groundwater quality (S. He et al., 2022). Its strength lies in its ability to capture complex interactions and nonlinear relationships between variables without necessitating extensive data preprocessing or handling missing values explicitly. Moreover, Random Forest provides insights into feature importance, allowing researchers to identify which variables significantly influence the model's predictions.

However, the algorithm is also prone to challenges such as high computational cost, overfitting problems and undesirable performance on high-dimensional data.

LightGBM, short for Light Gradient Boosting Machine, is another advanced gradient boosting framework that, like XGBoost, focuses on efficiency, speed, and performance but with reduced memory usage and higher efficiency (Ke et al., 2017). LightGBM employs a novel tree-growing algorithm, which significantly accelerates the learning process and reduces memory consumption without compromising accuracy. While it excels in managing large-scale datasets, it also presents certain limitations, including the need for extensive data preprocessing and substantial data volume requirements.

XGBoost, which stands for eXtreme Gradient Boosting, is a highly efficient and versatile machine learning algorithm that has gained popularity among data scientists and researchers across various fields (T. Chen & Guestrin, 2016). At its core, XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. It is particularly lauded for its ability to handle large and complex datasets with remarkable accuracy. For environmental engineers and scientists, XGBoost offers a powerful tool for tackling a wide range of modelling tasks, such as forecasting air pollution indicators (B. Pan, 2018), predicting soil pH distribution (S. Chen et al., 2019), and assessing groundwater quality (Singha et al., 2021). Its ability to deal with non-linear relationships and interactions between variables makes it well-suited for the multifaceted nature of environmental data. Additionally, XGBoost includes features that handle missing data and prevent overfitting, making it a robust choice for real-world environmental data analysis projects.

A typical workflow of the XGBoost algorithm is: 1) traverse all features in the dataset and sort the instances by eigenvalues separately; 2) determine the split points for each feature by finding the point with the highest information gain of all possible split points; 3) construct the optimal tree structure by choosing the best split strategy for all the features. Equation (2.7) shows the calculation of information gain in XGBoost:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \tag{2.6}$$

In the equation, G and H are defined as the sum of the first and second derivatives, respectively, of all the samples in a node L or R. λ and γ are constants. The formula could be decomposed as the gain scores on the new left branch, on the new right branch, on the original node, and the additional leaf's regularization. The node split only if the gain is greater than zero.

Physics-Informed Neural Networks (PINNs) is a promising AI-based solution for scientific problems, including environmental modelling (Zou et al., 2023; Karniadakis et al., 2021). These networks integrate information from physics equations into machine learning processes during the design or training phase, potentially mitigating some common limitations in traditional transport models, such as equifinality (Beven, 1996) and stiffness (Asaro & Lubarda, 2006) to a certain extent. The application of PINNs in environmental and earth sciences signifies a progressive shift towards more interpretable and accurate predictive modelling techniques necessary for solving complex multi-physics problems inherent in these domains. By integrating deep learning with established physical laws, PINNs can enhance the capability of model learning and predicting physical dynamics with appreciable accuracy. For instance, in soil hydrothermal modelling, PINNs have

demonstrated the ability to efficiently couple soil moisture and temperature dynamics, utilizing one to improve predictions for the other, thereby reducing reliance on dense measurement datasets (Y. Wang et al., 2023). This capability is of immense value in environmental engineering and science, where data scarcity often limits the robustness of model predictions. Further, in subsurface flow problems, PINNs have been successfully employed to estimate hydraulic conductivity under both saturated and unsaturated conditions, leveraging partial differential equation constraints to enhance accuracy. Additionally, the innovative use of PINNs with monotonicity constraints presents a novel approach for estimating soil-specific characteristics necessary for modelling water movement through soils directly from volumetric water content measurements without necessitating initial and boundary conditions (Raissi et al., 2019). The application of PINNs in capturing the hidden kinetics of complex biological processes like sulfur-driven autotrophic denitrification indicates the versatility and potential of these networks in analyzing intricate dynamics and designing process control strategies that outperform existing models (Zou et al., 2023). While being a great alternative to the traditional physics-based models (Ghorbani et al., 2021), PINN applications in contaminant transport modelling still harbour room for enhancements. For instance, the typical practice is to construct separate neural networks for each individual transport process (e.g., separate neural networks for adsorption and solute transport) rather than the whole system, increasing computational demand and impacting the feasibility and practicality of utilizing PINNs over classic physics models (Bertels & Willems, 2023; Bandai & Ghezzehei, 2022; Tartakovsky et al., 2020). Furthermore, the lack of clarity surrounding many transport factors for emerging pollutants such as metformin adds to the confusion in determining the

necessary prior knowledge to be utilized. Thus, a comprehensive methodology that retains advantages of PINNs while efficiently extracting and representing prior knowledge—particularly in scenarios with limited data—is highly desirable. Transfer learning is a machine learning technique where a model developed for one task is repurposed or fine-tuned on a related task, leveraging pre-existing knowledge to improve learning in the second scenario (S. J. Pan & Yang, 2010). It is particularly valuable in situations where labeled data for the second task is scarce or expensive to obtain. By transferring learned features, representations, and weights from a source model trained on ample data to a target model for a different but related task, the learning process is significantly accelerated and enhanced. This approach reduces both the time and resources needed for model training from scratch, making it an effective strategy for improving performance across a wide range of applications. It has shown promising applications across various disciplines, including environmental engineering and science, where the challenges of sparse data, data imbalances, and the high cost of data collection prevail. Transfer learning can significantly enhance the efficiency and accuracy of environmental predictions even with limited or imbalanced dataset availability by enabling the utilization of pre-trained models on new but related tasks. For instance, in predicting geogenic contaminated groundwaters, a Siamese Network-Based Transfer Learning model addressed the issue of insufficient groundwater quality data and class imbalances, achieving higher sensitivity and specificity in predicting hazardous substances' presence in groundwater compared to benchmark models (Cao et al., 2022). Similarly, the application for estimating dairy methane emissions from aerial imagery with transfer learning approaches showed a strong correlation with human visual inspections, offering a cost-effective alternative to traditional labour-

intensive methods (Jeong et al., 2022). In agriculture, transfer learning facilitated crop classification in regions with a shortage of training samples by leveraging crops' similar temporal growth patterns across different global regions (Hao et al., 2020). In addressing environmental concerns such as plastic pollution in soil, transfer learning methods have proven efficient in evaluating pollution levels across distinct soil regions using Near-infrared sensors, surpassing the performance of conventional multivariate analysis methods (Qiu et al., 2020). Predicting dynamic riverine nitrogen export in unmonitored watersheds has also benefited from transfer learning, leveraging insights from data-rich regions (Xiong et al., 2022). However, despite its advantages, transfer learning comes with limitations. The effectiveness of transferring knowledge largely depends on the relevance and similarity between the source and target tasks; if the tasks are too dissimilar, the transferred knowledge may not be beneficial and can even degrade the model's performance. Additionally, fine-tuning a pre-trained model on a new task requires careful adjustment of parameters to avoid overfitting, especially when the target dataset is small. Lastly, determining the optimal level of transfer and customization for the target task often involves trial and error, which can be time-consuming and require domain expertise.

2.4 Summary

This chapter has explored the existing methodologies, applications, and challenges within the realm of environmental data analysis, underlining its significance in addressing multifaceted environmental challenges. It specifically examined the applications and limitations of various techniques, including traditional environmental modelling tools, causal inference, machine learning, Physics-Informed Neural Networks (PINN), and transfer learning, providing a comprehensive overview of the landscape of environmental

data analysis. Despite remarkable advancements and the expanding versatility of environmental data analysis, the field encounters distinct challenges that necessitate ongoing research endeavors and methodological refinements. A principal concern is the evident lack of a comprehensive causal inference framework specifically crafted for intricate environmental datasets. This limitation constrains the depth of insight that can be derived regarding the causal relationships underlying observed patterns. Moreover, the application and development of physics-based models, despite their foundational importance, grapple with inherent limitations such as model stiffness and equifinality. These challenges underscore the necessity for innovative solutions that can enhance model adaptability and reliability, particularly in the context of new and emerging environmental threats. The escalation of such environmental threats brings to attention another critical challenge: the scarcity of data characterizing emerging environmental issues. This data gap hampers the ability to formulate timely and effective responses to novel pollutants and environmental changes, emphasizing the need for advanced data acquisition and analysis methodologies.

CHAPTER 3

MACHINE LEARNING-AIDED CAUSAL INFERENCE FRAMEWORK FOR ENVIRONMENTAL DATA ANALYSIS: A COVID-19 CASE STUDY[2]

[2] This chapter is based on and expanded from the following paper:
Kang, Q., Song, X., Xin, X., Chen, B., Chen, Y., Ye, X., & Zhang, B. (2021). Machine learning-aided causal inference framework for environmental data analysis: a COVID-19 case study. Environmental Science & Technology, 55(19), 13400-13410.
Roles: I conceived the study with the input from Song, X., Xin, X., Ye, X., Zhang, B. and Chen, B. I, Song, X., and Xin, X., contributed to data collection and processing. I, together with Chen, Y., developed the program script and conducted the data analysis. I wrote the first version of the manuscript, Zhang, B., Chen, Y. and Chen, B. contributed to subsequent revisions.

3.1 Introduction

After 12 months of the first COVID-19 case report in Wuhan, China (N. Zhu et al., 2020), a new SARS-CoV-2 variant was identified by the United Kingdom authorities on December 19, 2020 (Kirby, 2021). Two months later, the new variant with potential higher transmissibility and fatality (NERVTAG, 2021) has been found in ten Canadian provinces (Thompson, 2021) as well as in the United States and other 91 countries (O'Toole et al., 2021). As of June 30, 2021, multiple SARS-CoV-2 variants are circulating globally (Walensky et al., 2021), and the COVID-19 pandemic has claimed 3.93 million lives (Dong et al., 2020). The urgency of suppressing the COVID-19 pandemic has never been greater. Although SARS-CoV-2 can only be viable in aerosol for a limited period (3 - 16 hours; Fears et al., 2020; van Doremalen et al., 2020), COVID-19 was still reported to be capable of transmitting through the dissemination of suspended infectious aerosols (Bourouiba, 2020; Mittal et al., 2020; World Health Organization, 2020) in addition to unprotected contact with infectious individuals (J. F.-W. Chan et al., 2020; C. Huang et al., 2020) and fomite (contaminated surface; Chia et al., 2020; Guo et al., 2020; van Doremalen et al., 2020). Thus, as an effort to tackle the pandemic, the scientific community is examining factors associated with the pandemic, including environmental conditions such as meteorological factors and air pollution. As a result, correlations of air pollution and meteorological factors with COVID-19 severity have been reported worldwide (Accarino et al., 2021; Adams, 2020; Andree, 2020; Bashir, Ma, Bilal, Komal, Bashir, Farooq, et al., 2020; Carleton et al., 2021; Coccia, 2021c, 2021a; Haque & Rahman, 2020; Kulkarni et al., 2020; Y. Ma et al., 2020; Rahman et al., 2020; Rosario et al., 2020; Sarkodie & Owusu, 2020; X. Zhang et al., 2021). Those reported links lead to the speculation that some causal

mechanisms may exist behind the associations. For instance, low wind speed may promote the suspension of infectious particles (Coccia, 2020a, 2021b); exposure to air pollution may compromise people's immune systems and further induce a higher infection rate (Kutter et al., 2021; Srivastava, 2021; Tian et al., 2021).

Though no consensus has been reached (Islam et al., 2021; Qu et al., 2020; Sunyer et al., 2021), researchers deployed various approaches on several types of observational data, searching for clues to the causal links' existence. However, some issues are emerging while the research is becoming increasingly in-depth. The first issue is the confusion between correlation and causation (Holland, 1986). Due to ambiguous hypotheses and similarities between the two concepts, misidentifying the correlations as causalities is common. Another issue is the inappropriate use of conventional methods without the support of prior knowledge, which was constantly being overlooked in the existing studies. Those methods include time series analysis such as Granger causal test (Damette & Goutte, 2020; Delnevo et al., 2020; Mele & Magazzino, 2020) and machine learning models. Besides, some essential confounders, such as social-economical factors and inbound traffic flows from the pandemic epicentre (Bates et al., 2020; Coccia, 2020b; Pearl, 2000; Varian, 2016), were commonly omitted in the existing studies. Many spurious correlations could emerge due to such omission (Imbens & Rubin, 2015). Finally, among all the studies that attempted to estimate the causal effects quantitatively, few incorporated methods to refute the relationships or falsify the assumptions. The step is quite essential, especially when the ground truth of the causal links is unknown (Sharma & Kiciman, 2020). The issues above are not isolated but are commonly seen in environmental studies when a causal question has to be answered based on observational data without the aid of randomized experiments

(Rubin, 1974), such as in policy impact evaluation and climate change attribution (Forster et al., 2020; J.-Y. Liu et al., 2021). Thus, environmental studies can greatly benefit from a new framework for causal inference.

Thanks to the growing research on causal inference in the statistics and artificial intelligence field during the past few years (Butcher et al., 2021; Glymour et al., 2019; Prosperi et al., 2020), some novel and effective methods were born and thrived from the rich discussions, enabling us to develop a new causal framework with the desired features. To build such a framework which can conduct causal reasoning from observational data, among the most discussed methods, the SCM, one of the most established causal inference methods (Pearl, 2000) as the causal engine, was selected. The method has the following characteristics: 1) It uses prior knowledge regarding the data generating process as an input. 2) Intervention (i.e., purposely modify the condition to observe the response of the result) is a supported action in SCM in the form of do-calculus. The two features enabled SCM to perform causal reasoning from observational data. On the other hand, since the framework needs to be resilient to some common characteristics in environmental datasets such as frequent outliers, non-normal distribution and limited sample size (Ye et al., 2019), some functional components were also embedded to ensure the applicability and adaptivity of the framework. These components include: (a) a backup prior-knowledge extractor in case the prior knowledge is limited or not accessible; (b) a feature selection component, which can significantly reduce the computational time while acquiring data insights for the causal reasoning and (c) a refutation module that can test the proposed causal relation's robustness, which is especially helpful when the relationship is unconfirmed.

This chapter aims to propose a causal inference framework and to investigate the potential causal relationships between COVID-19 severity and environmental factors, including six air pollution indicators and four meteorological factors in 166 Chinese cities. The social-economic diversity among these cities makes China an ideal study area for investigating the causal relationships under multiple socio-economical scenarios (Huang, 2010). The study attempts to provide evidence for causal inference about environmental factors and COVID-19 severity, to support the decision-making process for global and regional pandemic countermeasures in the current phase of the COVID-19 pandemic and to establish an applicable and robust causal recovery framework for the environmental engineering and science community.

3.2 Materials and Methods

3.2.1 Framework Design, Study Area and Data sources

The workflow of the causal inference framework in this study is illustrated in Figure 3.1. This framework is suitable for environmental causal reasoning problems under different socio-economic conditions. During the data processing phase in the framework, socio-economic data will be used to generate clusters of different administrative units (i.e., countries, provinces, cities, etc.) with similar socio-economic conditions. Time series data from each administrative unit will be assigned to corresponding clusters. When the trends in the target time series are obvious, need-based time segmentation can be further applied. In that case, a time series segment (e.g., P_1 to P_n in Figure 3.1) will become the smallest unit for further analysis. Each segmentation will be analyzed by the machine learning module, followed by the causal inference module. Models for each unit will be trained by a selected machine learning algorithm then interpreted by multiple metrics for feature

selection. The interpretation can also support causal relationship identification in later procedures. Data will be input along with a DAG in the causal inference module. If no graph can be provided due to limited knowledge, a backup method can be called to generate a quasi-causal relationship graph as the DAG input. After quantitative estimation, the potential causal relationships will undergo two refutation processes as a robustness check.

This study investigated the causal relationship between environmental factors and COVID-19 severity. One hundred sixty-six key air quality monitoring cities recognized by the Ministry of Environmental Protection of China were selected as the study area due to their representativeness in their corresponding regions as well as their complete COVID-19 case and environmental monitoring data. Comparing to the original 168 key monitoring cities, Wuhan (the epicentre) was excluded and Dongying (had zero cases during the first wave pandemic) for the study. The socio-economic data were from each city's Statistical Bulletin/Yearbook or directly acquired from city-level Civil Affairs Bureaus. Most environmental data were acquired from the China National Environmental Monitoring Center. The COVID-19 related data were obtained from the Chinese Center for Disease Control and Prevention. The numbers of inbound travellers from Wuhan and the degree of activeness in each city were calculated based on Baidu Location-based Service (LBS) data.

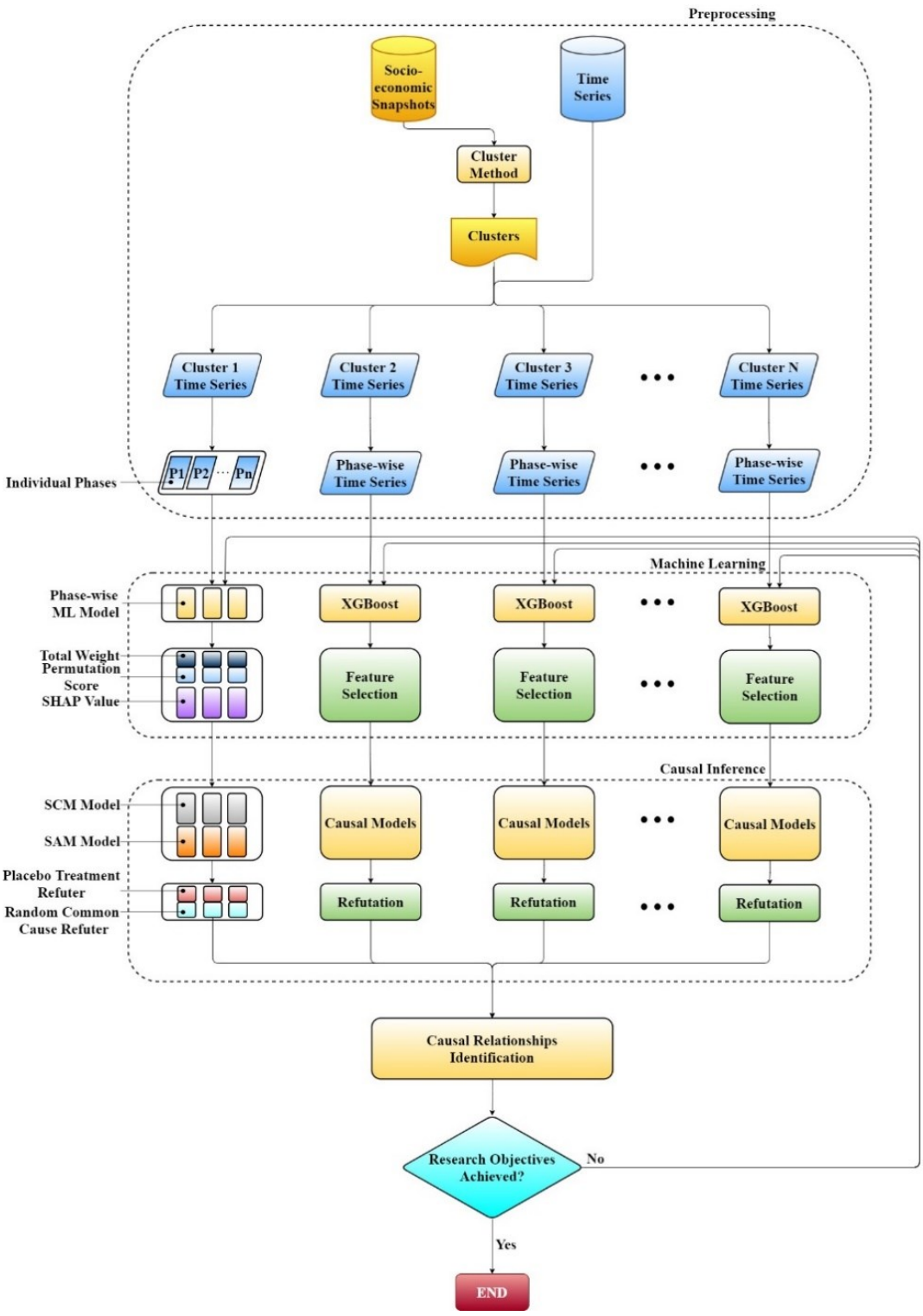


Figure 3.1 A schematic diagram of the causal framework. Note: "Cluster 1" indicates individual methods/parts in the component by making all the components transparent.

3.2.2 Measures of Variables and Data Processing

Two datasets, the "snapshot" dataset and the time-series dataset were composed and investigated. The "snapshot" dataset is a cross-sectional dataset consisting of all the 166 cities' socio-economic profiles before the 2020 Spring Festival. The features include:

- The inhabited populations (thousand people)
- Population density (people per km^2)
- Area of the cities (km^2)
- Total gross domestic products (GDP, in billion USD)
- GDP by sectors (primary, secondary, tertiary, in billion USD), and corresponding percentages
- GDP per capita (thousand USD)
- Elderly population percentage (over 60 years old)

  This feature was added since senior citizens are vulnerable to COVID-19 due to their fragile immune systems (Rothan & Byrareddy, 2020).

- Numbers of hospital beds, medical doctors, and nurses per thousand population

  The public healthcare development indexes were added, considering that the COVID-19 patients need timely and intensive care.

- Transient population flow from Wuhan (thousand people)

The 15-day accumulative inbound travellers from Wuhan to all the 166 selected cities before the pandemic outbreak were estimated according to Intracity Migration Index (IMI), a dataset developed based on Baidu's Location-Based Service. The dataset has also served the same purpose in previous related studies (Bao & Zhang, 2020).

- Average degree of activeness before the outbreak

  Based on the IMI data mentioned above, the degree of activeness in each city from January 10 to January 23, 2020, was used to calculate the average value.

The snapshot dataset was served as the basis for the following clustering process. A correlation heatmap of the snapshot dataset was given in Figure 3.2.

The time-series dataset comprises 13 time series for each selected city during the first wave of the pandemic. The 76-day lockdown period of Wuhan was selected as the time span for all the time series, which started from January 22 to April 8, 2020. The time series include:

- Six air pollutants' concentrations (PM2.5, PM10, SO_2, NO_2, and O_3 in $\mu g/m^3$, CO in mg/m^3)
- Average air temperature (TEMP, in ℃)
- Relative humidity (HMD, in percentage)
- Atmospheric pressure (PRES, in hpa)
- Wind speed (WSPD, in m/s)
- Daily degree of activeness (ACTV)

There are also two other features in the dataset:

- Daily new confirmed COVID-19 cases (CASES)

  In contrast to the moving average method, confirmed cases time series were processed with a three-day moving sum strategy for a more intuitive analysis process.

- Elapsed days (DAYS)

  Counted from the first day with a confirmed COVID-19 case in each city.

Three-day moving average to the above time series was applied to reduce the random noises in the dataset while focusing on the potential short-term effects (Jing et al., 2018; P. Li et al., 2013). In the time-series dataset, each feature's mean values in the corresponding city were used to impute a small portion (~0.23%) of missing values. A statistic description of two datasets and each feature's corresponding data source have been listed in Table 3.1.

Table 3.1 Features and Data Sources

Feature and Unit (if applicable)	Mean	Standard Deviation	Min	Max	Source
Snapshot Dataset					
Population (Thousand People)	5,624.67	4,029.80	720.96	31,243.20	B
City area (km^2)	11,733.64	9,080.77	1,459.00	82,402.00	Y
Population density (People per km^2)	652.19	694.77	24.31	6,729.49	E
GDP (Billion USD)	66.02	81.53	5.13	552.18	B
Primary sector (Billion USD)	3.54	2.51	0.17	21.87	B
Secondary sector (Billion USD)	25.80	26.94	1.89	151.89	B
Tertiary sector (Billion USD)	36.68	57.05	2.85	427.53	B
Primary sector percentage (%)	8.42	5.09	0.09	23.08	B
Secondary sector percentage (%)	41.33	7.58	16.16	60.00	B
Tertiary sector percentage (%)	50.25	8.08	33.54	83.52	B
GDP per capita (Thousand USD)	10.52	5.29	4.01	29.05	E
Elderly population percentage (%)	19.50	4.50	4.92	32.20	B, Y
Hospital beds per thousand people	6.22	1.22	3.82	9.67	B
Registered medical doctors per thousand people	2.81	0.76	1.32	5.76	B
Registered nurses per thousand people	3.19	1.01	1.27	6.71	B, Y
Travellers from Wuhan (Thousand People)	23.98	85.02	0.00	691.87	Baidu LBS
Wuhan travellers per thousand population	6.01	23.303	0.00	187.25	E

Average degree of activeness	5.36	0.64	2.98	7.08	Baidu LBS
Timeseries Dataset					
PM2.5 ($\mu g/m^3$)	46.67	31.34	3.67	349.00	CNEMC
PM10 ($\mu g/m^3$)	70.27	38.73	6.33	378.00	NEMC
SO_2 ($\mu g/m^3$)	10.33	7.415	1.67	92.00	NEMC
CO (mg/m^3)	0.81	0.35	0.20	4.50	NEMC
NO_2($\mu g/m^3$)	25.26	11.17	2.67	87.00	NEMC
O_3 ($\mu g/m^3$)	83.82	22.06	5.00	166.67	NEMC
Relative humidity (%)	71.23	18.20	8.00	100.00	BIN
Atmospheric pressure (hpa)	991.77	50.30	644.33	1,035.33	BIN
Wind speed (m/s)	2.23	1.31	0.10	11.47	BIN
Average air temperature	8.98	6.34	-22.00	27.68	BIN
Degree of activeness	3.59	1.34	0.31	8.81	Baidu LBS
New confirmed cases	6.17	34.26	0.00	1,021.00	CCDC
Morbidity rate	0.02	0.11	0.00	3.21	E

Note: “B” “Y” “E” in the Source columns indicate “Bulletin” “Yearbook” “Engineered Feature” respectively; “CCDC” is the abbreviation of the Chinese Center for Disease Control and Prevention; Percentage of Elderly and Registered Nurses per Thousand People were collected from multiple sources, including each city’s 2019 Statistical Bulletin, the 2018 Statistical Yearbook, or directly acquired from City-level Civil Affairs Bureau, depending on the availability of the data; The air pollution data was provided by the China National Environmental Monitoring Center, while the meteorological data was collected from an Application Programming Interface (API) provided by BINSTD, a data trading company; Travellers from Wuhan and Degree of Activeness in each city are calculated based on Baidu Location-based Service(LBS) data.

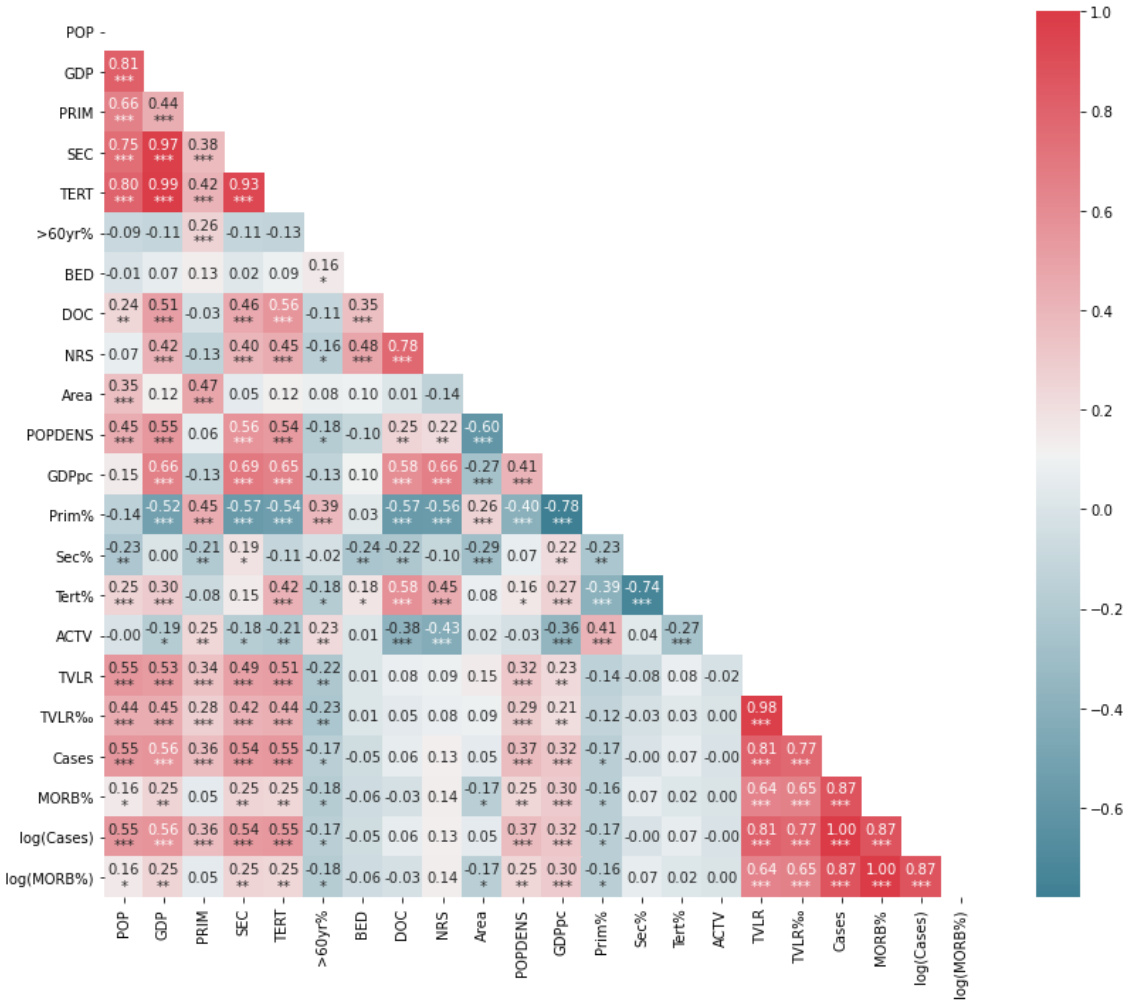


Figure 3.2 Spearman correlation heatmap of the snapshot dataset with statistical significance. Note: * indicates - $P \leq 0.05$, ** indicates - $P \leq 0.01$, *** indicates - $P \leq 0.001$. POP: Population; PRIM/SEC/TERT: Primary, secondary, tertiary sector of GDP; >60yr%: Elderly population percentage; BED/DOC/NRS: Hospital beds/registered medical doctors/registered nurses per thousand population; TVLR: Inbound travellers from Wuhan; TVLR‰: Wuhan travellers per thousand population; ACT: Average degree of activeness.

3.2.3 Models and Data Analysis

In this study, socio-economical factors should be considered essential since they are decisive for the human activity patterns and many other pandemic critical factors in different cities (Coccia, 2021e). Thus, the selected 166 cities were clustered based on the mentioned snapshot dataset. Such clusters can provide insights into the pandemic severity under different conditions and avoid the possible "Simpson's paradox" (Blyth, 1972). Principal Component Analysis (PCA; Pearson, 1901) was selected as the dimensionality reduction technique to ensure a better clustering performance and resilience to the "curse of dimensionality." Since PCA is sensitive to the variance within the dataset, the snapshot dataset was standardized before the procedure to minimize the impact of different feature scales and variance (Bellman, 2015; X. Song et al., 2019; Xin, Huang, An, & Feng, 2019; Xin, Huang, An, Raina-Fulton, et al., 2019). After a series of experiments, it was noticeable that over 62% of the variance could be explained with only three PCs, which is sufficient for further analysis. Thus, the original 18-dimensional dataset was compressed to three-dimensional by selecting three principal components. The contribution of each feature to individual principal components was given in Figure A.1, along with the explained variance ratios. For city clustering, the time-proven k-means method was selected due to its effectiveness and efficiency (Steinhaus, 1956). The "elbow method" (Thorndike, 1953) was used to determine the numbers of the cluster. Three was selected as the cluster number based on the result of the elbow method, which was given in Figure A.2. Due to the assumption that each factor may have different behaviours during different pandemic phases (Coccia, 2021d), the cluster-wise time series were further divided into two segments by specific demarcation points. The splitting dates were February 3, 2020, for Cluster 1

and February 6, 2020, for Cluster 2 and 3, corresponding to each cluster's pandemic spreading and post-peak phases. The corresponding trends can also be observed in Figure A.3(a). Hereafter, nine sub-datasets were generated from the time-series dataset, based on the pandemic development perspective (overall, spreading phase, post-peak phase) and three city clusters. Each sub-dataset will be further analyzed by both causal inference models and machine learning algorithms.

XGBoost algorithm was selected for feature selection and knowledge extraction. Models were trained with the aid of k-fold cross-validation. The cross-validation method splits an existing dataset into k number of folds, where each fold will be used as a testing set against the rest of the data. In this way, the impact of overfitting or sampling bias can be minimized. In the study, parameter k was set to 5 based on the number of instances (700 - 8500) in nine sub-datasets. After the training process, two feature importance evaluation metrics, total gain and permutation score, were used to interpret each trained model. The total gain in an XGBoost model is the product of a feature's Gain score and the frequency of the feature being used for node splitting when constructing the model. Permutation score is another useful metric defined as the decrease of the model performance when a single feature is randomly shuffled (Breiman, 2001). One common shortfall of the two metrics is their weakness in identifying if a features' contribution is positive or negative. Thus, SHAP was introduced as another method for interpretation as it can indicate the feature contribution's direction (i.e., positive or negative), which enabled the researchers to select features of interest for further analysis (Lundberg et al., 2020).

For causal inference, constructing a graphical causal model in the form of a DAG is the first step of the SCM (Pearl & Mackenzie, 2018; Sharma & Kiciman, 2020). Each DAG

node represents a variable, and an arrow indicates a causal link, either an assumed or confirmed one. The graph allows users to explicitly introduce prior knowledge and untested assumptions about the data-generating process. Figure 3.3 shows the graphic causal model for this study. Proven causal relationships are given as blue arrows. Causal relationships among elapsed days, degree of activeness and COVID-19 cases were considered, as well as those between meteorological factors and the air pollutants. The transformation from NO_2 to O_3 (Fahey et al., 1986) and the interactions among air temperature, relative humidity, wind speed and atmospheric pressure (Pearce et al., 2011) were also taken into account. Unproven causal links included potential causal relationships between the COVID-19 cases and different environmental factors. After creating the DAG, the Average Treatment Effects (ATE; Heckman, 1976, 1978) of the potential causal relationships could be estimated. The default incorporated algorithm is a linear estimator. To capture non-linear causal effects, the DMLOrthoForest (Microsoft Research, 2020) method was selected to provide non-linear estimations. The linear estimator has been preserved as a complement to the machine learning-based estimator. A more detailed introduction to the do-calculus and SCM can be found in the Literature Review. Note that causal estimation was conducted on a normalized copy of the dataset for enabling the comparison between different features.

Structural Agnostic Modelling (SAM) was deployed as the backup knowledge extractor. The neural-network-based algorithm has been proven robust in recovering non-linear causal relationships between continuous variables with a superb performance (Kalainathan et al., 2020). In this case study, though the DAG was constructed, SAM was used to generate a weighted adjacency matrix from the dataset, which is given in Table A.1. Each

weight in the matrix represents the corresponding causal relationship's strength. The matrix can be used to support the final decision-making about the causal relationship from another perspective.

In order to test the robustness of an assumed causal relationship, two refutation methods, Adding Random Common Cause (RCC) and Placebo Treatment (PT), were selected to test the robustness of each causal relationship. RCC adds an independent random variable as a common cause to the dataset, and PT replaces the chosen treatment variable's value with some independent random values. For a robust relationship, its estimated effect is expected to remain stable under the RCC refutation test. On the contrary, effects estimated under the PT test should be zero instead of the original value (Sharma & Kiciman, 2020). Based on the two refutation methods, a four-level robust check criterion was set in the case study to ensure the robustness of a causal estimation. Firstly, an estimate must pass both refutation tests, PT and RCC, to be considered. Being more specifically, the estimates under the RCC test should be within 10% variance of the original value, which is the first level. Then another three tolerance thresholds (i.e., the maximum allowed variation of an estimate) will be set to evaluate the considered estimations under the RCC test. In this study, the four levels were 10% (the initial threshold), 5%, 1% and 5‰, indicating an increasingly strict criterion. A potential causal relationship should pass the 5‰ threshold to be considered robust enough.

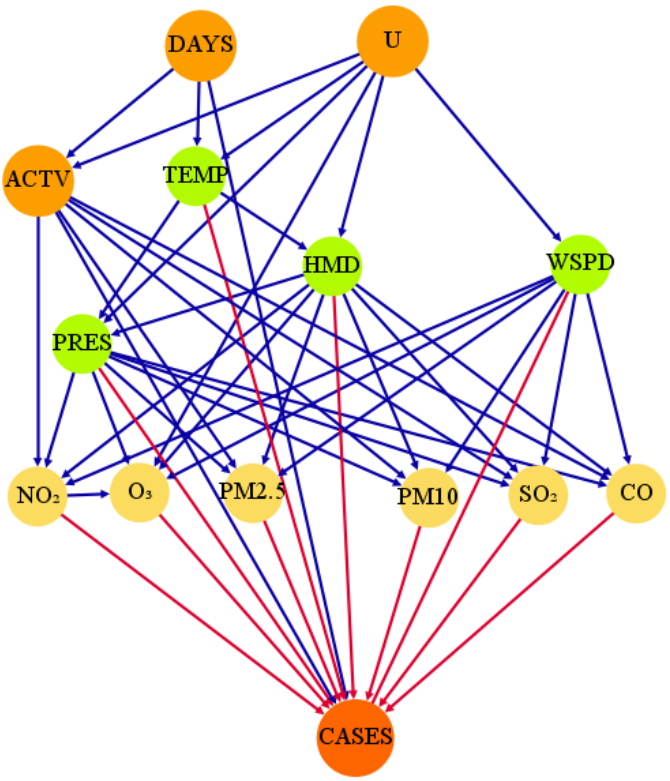


Figure 3.3 The causal relationships among environmental factors and COVID-19 cases. All proven causal links are given as blue arrows, and unproven causal relationships are marked by red arrows. Note: ACTV - Daily degree of activeness; DAYS - Elapsed days; HMD – Relative humidity; PRES - Atmospheric pressure; TEMP - Average air temperature; U – Unobserved confounders; WSPD – Wind speed.

The causal effect estimation and refutation were achieved based on the DoWhy package: a Python package specialized in providing a causal inference interface. The RCC and PT algorithms used in the framework can be found within the package (Sharma & Kiciman, 2020). The DMLOrthoForest algorithm and the SAM algorithm implemented in this study can be found in EconML (Microsoft Research, 2020) and CausalDiscoveryToolbox (Kalainathan & Goudet, 2019), respectively. A framework benchmark that applied SCM and SAM on three public datasets with known ground truth was given in the Appendix A.1 as a robustness check for the proposed framework. Meanwhile, some additional machine learning experiments were conducted during the study as an initial exploration, with their details listed below.

Removing the elapsed days feature. The elapsed days feature was removed in this experiment and trained the XGBoost models on the rest of the data. Feature importance was given in Figure 3.6. Under the setting, the air temperature became one of the top contributors among all clusters due to its high collinearity with elapsed time. It is reasonable to believe that introducing elapsed days can weaken spurious correlations originated from any other highly time-correlated features such as air temperature, and the feature importance for this experiment is given in Figure A.4.

Seven-day moving average. Instead of the original three-day moving average strategy, a seven-day moving average on the time series dataset was applied for machine learning and SCM analysis. No significant changes were observed. The results can be found in the GitHub repository of the study (https://github.com/kangqiao-ctrl/EnvCausal/tree/main/additional_experiments/7-day-moving-average).

Targeting cases per capita. In this experiment, the daily new cases per capita was used as the machine learning regression target instead of the absolute case number. No significant difference was observed in the result. It might because that the cities have already been clustered based on their socio-economic status include population. The results can be found in the GitHub repository of the study (https://github.com/kangqiao-ctrl/EnvCausal/tree/main/additional_experiments/morbidity_target).

3.3 Results

Three city clusters are presented in Figure 3.4 with the geographical locations of all the selected cities (n=166). The distribution of all cities in different clusters in the Principal Component Space is given in the Appendix A as Figure A.5. A full city list of all three clusters is available in Table 3.2. In summary, Cluster 1 (n=7) comprised megacities with advantages in many socio-economic aspects. Cities in Cluster 2 (n=40) are mostly provincial capitals and other major cities, and the majority of Cluster 3 (n=119) are ordinary urbanized cities. The average feature values in different city clusters are given in Table 3.3. No significant difference could be found in the elderly population percentages (16.54% to 20.62%) and healthcare development indexes (beds per thousand: ~6.38, doctors per thousand: ~3.19, nurses per thousand: ~3.79) among the three clusters. The degrees of average activeness in the three clusters were also at the same level (~5.09) though a relatively higher activeness degree (5.57) can be observed in Cluster 3.

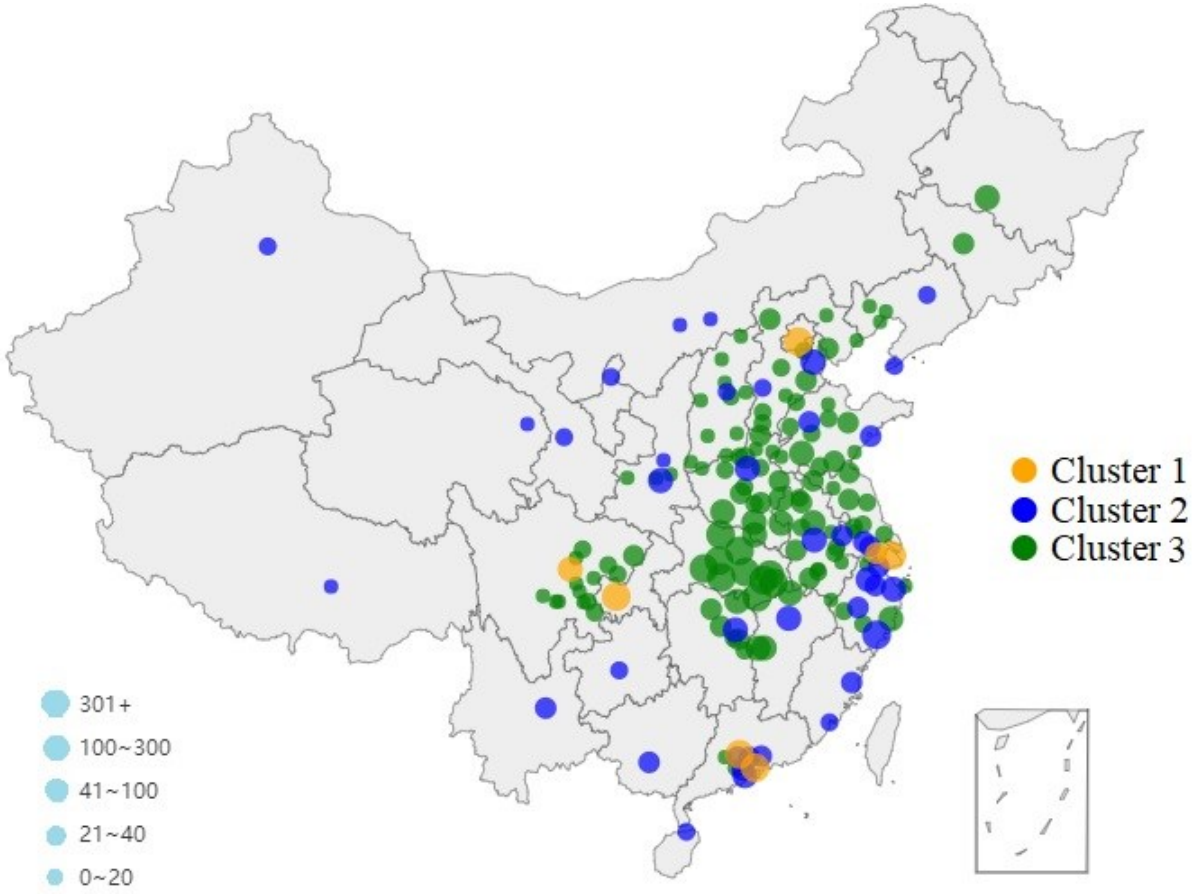


Figure 3.4 The selected cities' locations with the circle size indicating the total confirmed COVID-19 cases.

Table 3.2 Cities in different clusters

Cluster	City names
Cluster l (Megacities)	Beijing, Shanghai, Chongqing, Suzhou, Chengdu, Guangzhou, Shenzhen
Cluster 2 (Major Cities)	Shenyang, Dalian, Fuzhou, Xiamen, Nanning, Haikou, Guiyang, Kunming, Lhasa, Lanzhou, Xining, Yinchuan, Ürümqi, Tianjin, Shijiazhuang, Taiyuan, Jinan, Qingdao, Zhengzhou, Hohhot, Baotou, Nanjing, Wuxi, Changzhou, Hangzhou, Ningbo, Wenzhou, Shaoxing, Jiaxing, Jinhua, Hefei, Xi'an, Tongchuan, Nanchang, Changsha, Zhuhai, Foshan, Huizhou, Dongguan, Zhongshan
Cluster 3 (Common Cities)	Chaoyang, Jinzhou, Huludao, Changchun, Harbin, Tangshan, Qinhuangdao, Handan, Xingtai, Baoding, Zhangjiakou, Chengde, Cangzhou, Langfang, Hengshui, Datong, Shuozhou, Xinzhou, Yangquan, Changzhi, Jincheng, Lüliang, Jinzhong, Linfen, Yuncheng, Zibo, Zaozhuang, Weifang, Jining, Tai'an, Rizhao, Linyi, Dezhou, Liaocheng, Binzhou, Heze, Kaifeng, Pingdingshan, Anyang, Hebi, Xinxiang, Jiaozuo, Puyang, Xuchang, Luohe, Nanyang, Shangqiu, Xinyang, Zhoukou, Zhumadian, Luoyang, Sanmenxia, Xuzhou, Nantong, Lianyungang, Huai'an, Yancheng, Yangzhou, Zhenjiang, Taizhou, Suqian, Huzhou, Quzhou, Taizhou, Lishui, Zhoushan, Wuhu, Bengbu, Huainan, Ma'anshan, Huaibei, Tongling, Anqing, Huangshan, Fuyang, Suzhou, Chuzhou, Lu'an, Xuancheng, Chizhou, Bozhou, Xianyang, Baoji, Weinan, Zigong, Luzhou, Deyang, Mianyang, Suining, Neijiang, Leshan, Meishan, Yibin, Ya'an, Ziyang, Nanchong, Guan'gan, Dazhou, Xianning, Xiaogan, Huanggang, Huangshi, Ezhou, Xiangyang, Yichang, Jingmen, Jingzhou, Suizhou, Pingxiang, Xinyu, Yichun, Jiujiang, Zhuzhou, Xiangtan, Yueyang, Changde, Yiyang, Jiangmen, Zhaoqing

Table 3.3 Average feature values in different city clusters

Feature and Unit (if applicable)	Cluster 1	Cluster 2	Cluster 3
Population (Thousands)	19,019.47	6,266.86	4,620.88
City Area (km^2)	19,653.14	10,440.70	11,702.39
Population Density (People per km^2)	2,386.85	868.45	477.46
GDP (Billion USD)	380.27	97.44	36.97
Primary Sector (Billion USD)	5.82	2.92	3.61
Secondary Sector (Billion USD)	117.50	39.16	15.92
Tertiary Sector (Billion USD)	256.87	55.37	17.44
Primary Sector Percentage (%)	18.25	34.68	10.47
Secondary Sector Percentage (%)	32.57	39.55	42.45
Tertiary Sector Percentage (%)	65.58	57.00	47.08
GDP Per Capita (Thousand Yuan)	147.6	105.19	57.13
Elderly Population Percentage (%)	17.46	16.54	20.62
Hospital Beds per Thousand People	6.37	6.73	6.04
Registered Medical Doctors per Thousand People	3.49	3.57	2.51
Registered Nurses per Thousand People	4.30	4.32	2.74
Wuhan Travellers (Thousand People)	31.69	7.49	29.07
Wuhan Travellers per Thousand Citizens	15.93	10.72	7.93
Average Degree of Activeness	4.87	4.84	5.57

Figure 3.5 shows the feature importance and the SHAP values. Features with positive contributions in each trained machine learning model were highlighted. Both the total gain and the permutation score were normalized to a 0-1 range for easier comparison and visualization. The hyperparameters ranges for GridSearchCV, the final hyperparameter values and r^2 of each trained XGBoost model are given in Table 3.4. Note that as two baseline features, elapsed days and degree of activeness were designed to be never highlighted. For elapsed days, it is noticeable that the sign of its contribution varied in different sub-datasets. It was the dominating factor among all three phases in Cluster 1 with normalized feature importance above 0.60, the top contributor for the post-peak phase in Cluster 2 with feature importance of 0.35, and had lower feature importance varied from 0.05 to 0.15 for other periods in Cluster 2 and 3. Similar to elapsed days, the degree of activeness also dominated in one cluster, Cluster 3, with its feature importance maintained around 0.29. It reached the top in Cluster 2 from the overall perspective with feature importance of 0.23, had lower yet considerable feature importance (~0.15) in the other two Cluster 2 phases, and became less significant in Cluster 1.

Table 3.4 Final XGBoost model hyperparameters and R^2

	Cluster 1			Cluster 2			Cluster 3		
	Spreading	Post-peak	Overall	Spreading	Post-peak	Overall	Spreading	Post-peak	Overall
max_depth	4	4	2	4	3	3	4	5	5
min_child_weight	9	8	8	9	8	8	8	6	3
n_estimators	50	50	50	50	50	50	50	50	50
learning_rate	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
R^2	0.965	0.965	0.796	0.891	0.806	0.768	0.877	0.842	0.837

*The ranges of the hyperparamters used in GridSearchCV are given as below: max_depth [2,10]; min_child_weight [2,10]; n_estimators [25,50,75,100,150,200,250,300]; learning_rate [0.01,0.05,0.1,0.2,0.3]

From the interpretation results, positive contributions of air pollutants were generally low. Most air pollutants' feature importance was below 0.05, except for a few specific pollutants under particular settings. The exceptions included SO_2 in Cluster 1 spreading phase (0.10), CO from Cluster 3 overall perspective (0.06) and post-peak phase (0.14), as well as PM2.5 (0.07) from the overall perspective in Cluster 3. Meteorological factors had a higher potential contribution than air pollutants, which can be observed in Figure 3.5. All meteorological factors had been highlighted at least once. Atmospheric pressure was highlighted among all sub-datasets for nine times. Its feature importance was within a range of 0.08 to 0.20 in Cluster 2 and Cluster 3, whereas being less significant in Cluster 1 with its feature importance varied from 0.03 to 0.09. The air temperature was the second-most highlighted meteorological feature, which had been highlighted six times. When highlighted, its feature importance varies from 0.02 to 0.22. Relative humidity showed most of its observable potential contribution in Cluster 3, with a feature importance ranging from 0.03 to 0.10. Wind speed had some minor contribution less than 0.03 in Cluster 3.

Table 3.5 Refutation results of potential impactful environmental factors

City Cluster	Pandemic Phase	Feature	Relationship Type	ATE	Threshold		
					5%	1%	5‰
Cluster 1	Overall	PRES	Non-linear	0.045	F		
	Spreading	SO_2	Non-linear	0.476	F		
	Post-peak	HMD	Linear	0.051	P	P	F
Cluster 2	Overall	PM2.5	Non-linear	0.323	P	F	
		O_3	Non-linear	0.012	P	F	
		PRES	Non-linear	0.055	P	F	
	Spreading	PRES	Non-linear	0.118	P	F	
		TEMP	Non-linear	0.041	P	P	P
	Post-peak	PM2.5	Non-linear	0.290	P	F	
		HMD	Non-linear	0.046	F		
		PRES	Non-linear	0.066	P	F	
		TEMP	Non-linear	0.030	P	F	
Cluster 3	Overall	O_3	Non-linear	0.016	P	P	F
		HMD	Non-linear	0.018	P	F	
		PRES	Non-linear	0.008	F		
		WSPD	Non-linear	0.009	P		
	Spreading	PM10	Linear	0.079	P	P	F
		SO_2	Non-linear	0.057	P	F	
		HMD	Non-linear	0.046	F		
		TEMP	Non-linear	0.073	F		
		WSPD	Non-linear	0.005	P	F	
	Post-peak	PM10	Non-linear	0.035	F		
		CO	Non-linear	0.092	F		
		PRES	Non-linear	0.016	F		
		WSPD	Non-linear	0.022	P	F	

Note: Non-linear: Estimated by DMLOrthoForest; Linear: Estimated by linear estimator; ATE: Average Treatment Effect; Threshold: the maximum allowed variation of an estimate, used to evaluate the RCC refutation results; P: Pass; F: Fail. PRES: Atmospheric Pressure; HMD: Relative Humidity; TEMP: Air Temperature; WSPD: Wind Speed. All the estimates in the table passed the RCC test with a 10% threshold.

Table 3.5 shows the 25 potential causal relationships that passed the initial refutation with positive ATE, as well as their behaviours when facing lower RCC tolerances. Nine out of the twenty-five relationships were about air pollution indicators. The majority of the connections were non-linear with two linear exceptions: relative humidity (Cluster 1 post-peak phase) and PM10 (Cluster 3 spreading phase). The effect of reducing the RCC threshold was prominent. Decreasing the tolerance from 10% to 5% eliminated nine candidate relationships. Tolerance's dropping from 5% to 1% removed another 11 candidates. NO_2, one of the most reported air pollutants with a correlation between COVID-19 severity, did not pass the initial causal screening. As for SO_2 in Cluster 1, although it had positive contributions in Cluster 1 machine learning models and passed the initial round of the refutation test, it did not survive the first tolerance drop. When the tolerance dropped to 5‰, only one potential causal relationship survived: air temperature in Cluster 2 spreading phase with a causal effect of 0.041.

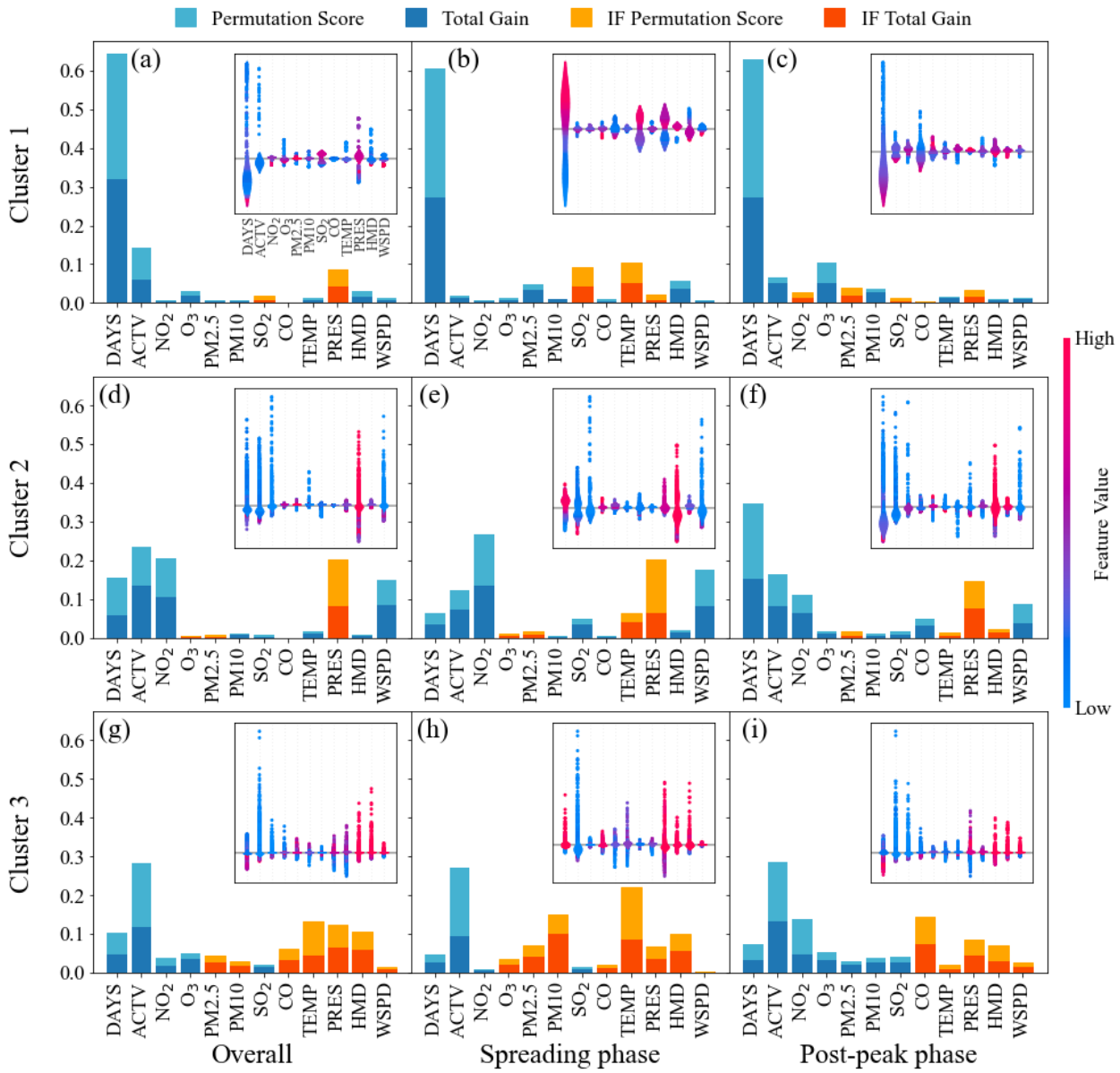


Figure 3.5 Feature importance and SHAP value of features in machine learning models. Features with positive SHAP values (subplots) are highlighted with orange and red in the feature importance plot. In the SHAP value subplots blue and red indicates lower and higher feature values, respectively. Instance points above the grey zero axis indicate positive SHAP values and vice versa. SHAP value features are in the same order as in subplot (a).

3.4 Discussions

It is noticeable that the majority of the pollutants' contributions in the machine learning models were negative. This characteristic can be easily observed from NO_2, which had significant negative contributions (0.1 ~ 0.2) and was considerably correlated with the degree of activeness (ρ=0.47) in Cluster 2. Meanwhile, the negative contributions were reflected in the estimated effects from the SCM as well: all of NO_2's non-linear ATE values were negative. From the results, NO_2 is more likely to be an indicator of human activity in selected Chinese cities rather than a causal factor to COVID-19 cases. O_3, another air pollutant that may compromise the human respiratory system (W. Gao et al., 2017; A et al., 2008), also had some negative causal effects (four out of nine), indicating it was unlikely to worsen the pandemic, especially considering that none of its relationships passed the final refutation. It can be assumed that those negative contributions were due to the connections between air pollution and human activity: the primary sources of NO_2 and O_3 in China are anthropogenic activities, mostly from industrial and mobile sources (LIU et al., 2020; Xue et al., 2014). Implement and lift the lockdown policies might further influence the fluctuation of the pollutants' concentrations (Diao et al., 2021; Shen et al., 2021; K. Xu et al., 2020). Note that the genuine causal relationships among lockdown implementation, human activities and air pollution do not guarantee observable correlations. A visualization of NO_2 and O_3's trends was given in Figure A3.3(d) and (e).

Among all the relationships about PM2.5, PM10, SO_2 and CO, PM10 in Cluster 3 spreading phase was the only relationship that passed the 1% threshold refutation test with positive ATE values (0.079), though it did not pass the final refutation. As for PM2.5, both its relationships in Table 3.5 could not pass the 1% threshold refutation. The only CO-related causality failed the 5% threshold test in the Cluster 3 post-peak phase; SO_2 in Cluster 3 spreading phase failed the 1%

refutation. The results indicate the robustness of those causal relationships was not sufficient. The same deductions can be applied to the meteorological relationships, where most of them failed the second level of refutation (5%), implying their insignificance in the proposed causal problem. Though the majority of the relationships were refuted at the end, air temperature in the Cluster 2 spreading phase passed the final refutation with a causal effect of 0.041. Technically, the values indicate that 1 ℃ air temperature increase in Cluster 2 during the spreading phase will lead to approximately 0.183 new confirmed cases. However, its final RCC refutation variance was 0.00498, indicating that the temperature-case causal estimation almost failed the final refutation test (5‰ threshold). Based on all the results, though a specific causal relationship's existence cannot be completely ruled out, the discussed factors' causal effects on the COVID-19 severity are likely to be limited. Since the estimates reported in the study were by no means conclusions, but traces of evidence of the causal links. Thus, instead of drawing conclusions, it is more reasonable to deduce that the environmental factors were unlikely to exacerbate the COVID-19 pandemic in these Chinese cities from a short-term perspective.

3.5 Summary

In this chapter, we introduce a sophisticated causal inference framework that leverages SCM enhanced by machine learning techniques to scrutinize the potential causal links between environmental factors and COVID-19 severity across 166 Chinese cities. By incorporating prior knowledge and employing a comprehensive data processing strategy that includes city clustering and phase-wise analysis, socio-economic and temporal considerations were brought into the causal investigation. Utilizing this framework, this chapter study meticulously evaluates reported causal relationships between ten environmental variables (including NO_2, O_3, PM2.5, PM10, SO_2, CO, average air temperature, atmospheric pressure, relative humidity, and wind speed) and the severity of COVID-19, categorizing cities into three clusters based on socio-economic characteristics and analyzing time series data across different phases of the pandemic, and refuted the majority of these potential causal links (89 out of 90). This detailed investigation not only underscores the limited influence of environmental factors on the pandemic's severity but also showcases the framework's capability to address causal questions with observational data, thereby enriching environmental research and other disciplines. This chapter also demonstrated the high value and potential of the proposed framework in investigating causal problems with observational data in environmental or other fields.

CHAPTER 4

CAUSAL PRIOR-EMBEDDED PHYSICS-INFORMED NEURAL NETWORK AND A CASE STUDY ON METFORMIN TRANSPORT IN POROUS MEDIA[3]

[3] This chapter is based on and expanded from the following manuscript: **Kang, Q.,** Zhang, B., Cao, Y., Song, X., Ye, X., Li, X., Wu, H., Chen, Y. & Chen, B.* (2024). Causal Prior-Embedded Physics-Informed Neural Networks and a Case Study on Metformin Transport in Porous Media, *Accepted by Water Research.*
Roles: I conceived the study with the input from Song, X., Zhang, B. and Chen, B. I, Cao, Y., and Song, X., contributed to conducting the transport experiments and gathering data. I developed the program scripts for processing, analyzing and visualizing the data along with Ye, X., and Wu, H. Visualization and interpretation of the results were further enhanced by the contributions from Cao, Y., Song, X., and Li, X. The initial manuscript draft was written by me, with subsequent feedback provided by Zhang, B., Chen, Y. and Chen, B.

4.1 Introduction

As we transition into an era of data explosion, the field of environmental modelling is experiencing profound transformations accompanied by two challenges. The primary challenge involves the adoption of artificial intelligence (AI) within environmental modelling. While AI can offer superb performance and enhanced efficiency, traditional AI applications often incorporate minimal expert knowledge. This challenge can result in a deficiency in robustness and, in some cases, lead to misinterpretation of the model results (Rolnick et al., 2023; Zhong et al., 2021). The second challenge is the difficulty in effectively extracting, refining, or utilizing prior knowledge. This becomes particularly pronounced when dealing with limited information, such as in the context of emerging pollutants like Pharmaceutical and Personal Care Products (PPCPs). PPCPs are recognized as one of the largest groups of emerging pollutants due to their widespread occurrences, lack of sufficient regulation, and persistence in the environment (Schwartz et al., 2021; Wilkinson et al., 2022). Metformin, a commonly prescribed medication for type-2 diabetes, serves as a prime example of a PPCP. It has emerged as an environmental concern due to its reported endocrine-disruptive properties, ecological impacts, and the adverse health effects of its chlorination by-products (R. Zhang et al., 2021; Niemuth & Klaper, 2018; Briones et al., 2016; Niemuth et al., 2015). Its ubiquitous occurrences (ICPDR, 2020; Tao et al., 2018; Yao et al., 2018; Oldenkamp et al., 2018) in various environmental compartments, including in groundwater, soil and drinking water sources, highlights the potential risk that it may infiltrate potable water sources via soil-to-groundwater pathways (Y. He, Zhang, et al., 2022; Lesser et al., 2018; Tisler & Zwiener, 2018; Trautwein et al., 2014). The transport of metformin in the subsurface environment has garnered some

attention (Briones & Sarmah, 2019, 2018a; Lopez et al., 2015). However, it is challenging to yield a comprehensive, universally applicable understanding of metformin's transport dynamics with limited data. Thus, transport problems related to emerging pollutants, such as metformin, could greatly benefit from a data-driven framework that efficiently embeds constrained expert knowledge without compromising performance (Gibert et al., 2018).

The causal prior-embedded physics-informed neural network is becoming a promising approach. It combines data-driven methods with causal prior knowledge extracted from enhanced experiment data. The extracted causal information is then incorporated into a data-driven application using various advanced causal embedding techniques. The task of the data-driven application mirrors a classic transport model, such as predicting the breakthrough time of a solute in a specific porous medium condition. In the proposed method, the SCM, a graph-based causal inference method, was selected to extract causal knowledge from the experimental dataset (Butcher et al., 2021; Prosperi et al., 2020; Glymour et al., 2019; Pearl, 2000). The rationale behind utilizing SCM arises from the recognition that in contaminant transport processes, most variables are intricately and causally interconnected, signifying that a change in a cause variable will inevitably lead to changes in some others. To estimate their causal strengths offers a good approximation of the underlying physical processes. Successful estimations can be transformed into a weighted version of DAG, where each edge corresponds to causal strength. This weighted DAG is more computationally efficient when embedded in downstream data-driven applications than traditional transport models PDEs. Considering SCM might be well-suited for knowledge extraction in the context of contaminant transport (Um et al., 2019), it was selected for this framework.

The proposed method possesses an additional layer of interpretability by infusing *causal priors* derived from PDE-based models and requires fewer computational resources than typical PINN applications. Furthermore, its flexible and scalable architecture enables compatibility with various problem domains and data volumes. This positions it as a versatile tool for scientific data-driven applications. Additionally, the influence of various neural network hyperparameters within the context of causal embedding was closely examined. This helped us evaluate the robustness and broad applicability of our methodology in scientific contexts. The one-dimensional transport of metformin in sandy columns serves as a case study to validate our proposed method, to investigate its transport behaviour within the soil and groundwater compartments and enhance our understanding of the emerging pollutant. By proposing the causal physics-informed prior embedded neural network, this chapter seeks to strike a balance between data-driven methods and expert knowledge, thereby providing a robust and innovative solution in the field.

4.2 Materials and Methods

4.2.1 Framework Design

The schematic representation of our research framework is depicted in Figure 4.1. Experimental data acquired under various representative experimental configurations is subsequently fed into physics-based transport models such as Richards' equation to estimate plausible distributions of unmeasurable parameters $\hat{\omega}$ (for example, type-1 sorption fraction). The experimental data with the physics model-estimated parameters serves as the "seed data," which is a foundational dataset used to initialize and inform further modelling efforts (Ebert-Uphoff & Deng, 2017). Following this, a grid-search-based data augmentation is performed, with a series of constraints ensuring the resulting augmented dataset can represent the substance's transport in the specific transport media with limited distortion. This dataset is then divided for causal inference and neural network training. Simultaneously, a causal graph is constructed in the form of a DAG, incorporating most of the impactful variables based on prior knowledge of the system. This step initiates the SCM, a graph-based causal inference methodology that identifies causally impactful variables in a multivariate system. It allows for quantitative estimation of causal impact using various regression techniques. The estimated causal dynamics are referred to as the "*causal prior*." These causal priors are embedded into the model through two techniques: *causal weight initialization*, which is introduced in this study and an existing method, *causal prior regularization* (Kancheti et al., 2022), to enhance interpretability in data-driven models like multilayer neural networks. Ultimately, the causal-embedded neural networks are evaluated from two angles: model performance after embedding the causal prior and retained causal interpretability (hereafter referred to as "causal retention").

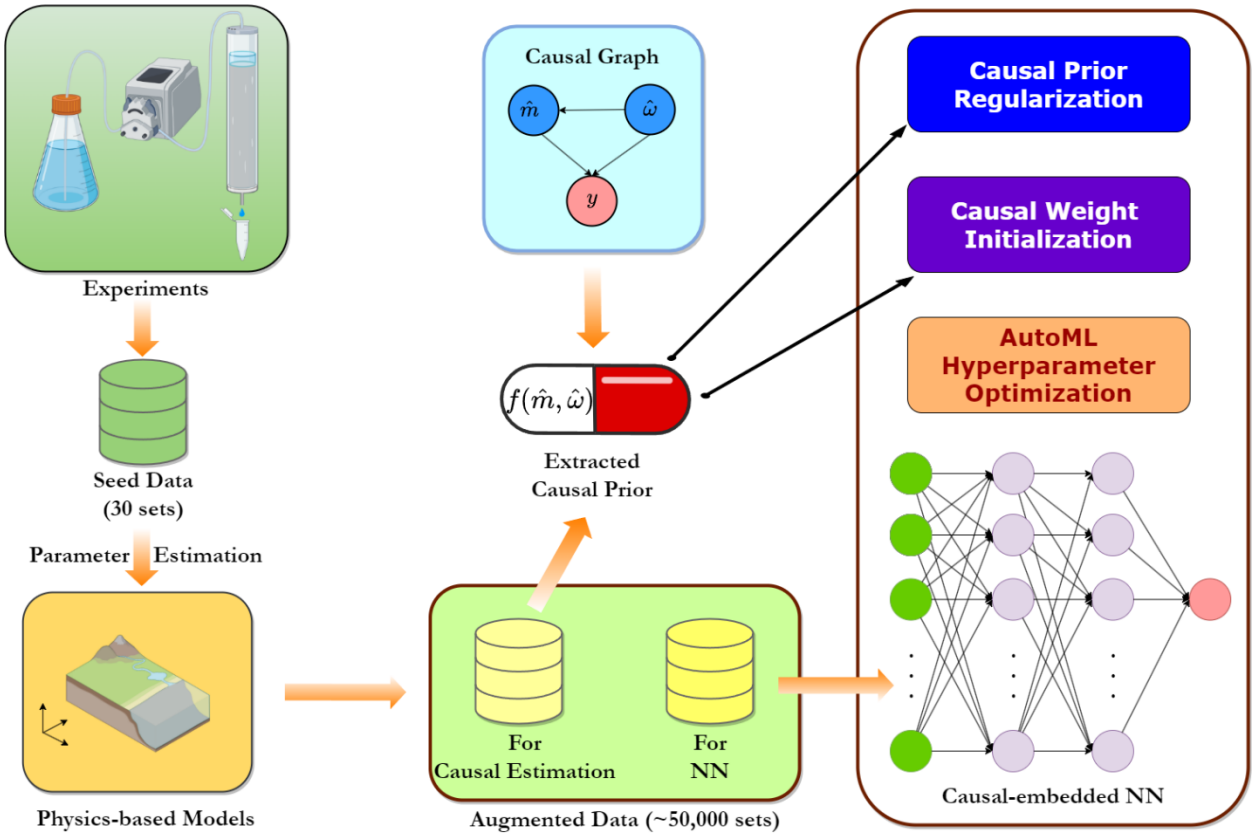


Figure 4.1 A schematics diagram of the proposed causal embedded physics-informed neural network. In the *causal prior* extracted from a substance transport process, $\hat{\mathrm{m}}$ and $\hat{\omega}$ represent measurable parameters (e.g., boundary conditions) and parameters that cannot be directly measured or need to be acquired from the curve fitting process, respectively, and $\mathrm{f}(\hat{\mathrm{m}}, \hat{\omega})$ is a function corresponding to the causal dynamics within the system, i.e., a function from all the physics meaningful factors within the system to the variable of our interest.

4.2.2 Transport of Metformin in the Sandy Media

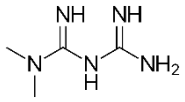
4.2.2.1 Experimental Materials

Commercial quartz sand (>98%) was used as the porous media for metformin transportation. The sand went through 20/40 (0.850 mm/0.425 mm), 40/80 (0.425 mm/0.180 mm), and 80/120 (0.180 mm/0.125 mm) meshes, and hereinafter referred to as *coarse*, *medium*, and *fine sand*, respectively. Another two types of sand, *"mixed sand (variant 1)"* and *"mixed sand (variant 2),"* were acquired from 2-mm meshed quartz sand without going through finer meshes. All the types of sand underwent acid (HCl) and alkaline (NaOH) solution wash and then rinsed with de-ionized water to ensure the removal of all the impurities (Rostvall et al., 2018). The grain size distributions of the sandy media were calculated by a particle size analyzer, and the grain size distributions for mixed sands (variants 1 and 2) are given in Table 4.1. Metformin hydrochloride (98%, manufactured by MP Biomedicals™) and sodium azide crystalline (NaN_3) were purchased from Fisher Scientific Company, Ontario, Canada. The physicochemical properties of metformin are given in Table 4.2.

Table 4.1 Grain size distribution (%) of different types of sand

Grain size (mm)	2-1	1-0.5	0.5-0.35	0.35-0.25	0.25-0.15	0.15-0.075	<0.075
Mixed Variant 1	2.48	21.79	24.86	9.54	24.52	11.71	5.1
Mixed Variant 2	2.41	12.56	15.43	9.05	32.34	19.66	8.55

Table 4.2 Metformin physiochemical properties and fate parameters

Parameter	Unit	Metformin
CAS Number	-	657-24-9
Structure		
Molecular Weight	g/mol	129.2
Water Solubility (25 °C)	g/L	300
Log K_{OW}	-	-4.9
Log K_{OC} (Soil)	-	3.05
Log K_{OC} (Activated Sludge)	-	39.2
Vapour Pressure	mm Hg	7.58×10^{-5}
K_H	$atm/m^3/M$	7.6×10^{-16}
pKa (Calculated)	-	10.0, 12.3
pKa (Experimental)	-	3.1, 13.8
Melting Point	℃	223-226
Boiling Point	℃	268.97
Aerobic biodegradation		Degradation after 28 days ≈ 0.6 %
Hydrolysis half-life	year	> 1
Photolysis half-life	day	28.3

Most of the information was summarized from Briones' metformin Global occurrence research (Briones et al., 2016), and metformin fate parameters were from AstraZeneca's report (AstraZeneca, 2020).

4.2.2.2 Experiment Settings

The 1-D column experiments were carried out to investigate the transport of metformin in sandy media. Acrylic columns with lengths of 80 cm, 60 cm and 24 cm with pre-drilled inlets and outlets were used to create 1-D flow spaces with different dimensions. The experiment settings were categorized based on three flow conditions: saturated and top-down (with ponded water); unsaturated and top-down (no ponded water); saturated and bottom-up, as illustrated in Figure 4.2 (a, b, c). Under settings (a) and (b), two column sizes were used: 80 cm in length and 8 cm in internal diameter, and 40 cm in length and 5 cm in internal diameter. These columns were filled with four types of sand: *medium, fine* and *mixed (variants 1 and 2)* to the designated depths of 60 cm for the longer columns and 27 cm for the shorter ones. In setting (c), designed to explore the effect of controlled flow rate and varying concentrations, smaller columns (24 cm in length and 2.4 cm in internal diameter) were entirely filled with *coarse sand*. The sand was compacted every 3 cm during filling to achieve a consistent density throughout the profiles. For each end of the column, four layers of gauze were placed to prevent clogging. To measure the residual water content of different column systems, the columns were moistened by pumping de-ionized water through them until saturation, followed by an overnight gravity drain to leave only residual water content. De-ionized water was then introduced into the columns to establish varying saturation and flow conditions. To get different saturation conditions for different scenarios, for setting (a), the inflow was adjusted until a ponded water of 3 cm was maintained; for setting (b), the inflow was balanced with outflow rates; and for setting (c), columns were fully saturated and the inflow was set to four predetermined levels (0.444,

0.666, 0.888, 1.11 ml/min) while the outflow rates were examined to ensure they were the same as the inflow rate. All flow rates were documented for further analysis.

A metformin stock solution with a concentration of 500 mg/L was prepared for subsequent column experiments, containing 0.02% sodium azide (NaN_3) to suppress any potential biodegradation process (MacQuarrie et al., 2001; Groffman et al., 1996).

4.2.2.3 Tracer and Metformin Column Breakthrough Experiments

As a non-reactive tracer, NaCl solution at 1,500 mg/L was introduced to all the columns to estimate transport parameters through curve fitting with groundwater models, after the flow and saturation conditions stabilized. The tracer experiment continued until the concentration in the collected samples was consistently nearly equal to the concentration in the influent. Following tracer testing and a subsequent rinse with de-ionized water, the inflow was switched to metformin solutions at various concentrations, and the experiment ended until the concentration in the collected samples was consistently nearly equal to the concentration in the influent. For settings (a) and (b), a fixed 10 mg/L concentration of metformin was used, while for setting (c), the following concentration was used: 2.5, 5, 10, 20, and 40 mg/L, respectively.

Two-milliliter samples from both the tracer experiment and the metformin transport experiment were collected at fixed time intervals (2, 4, 10, 30, 60 minutes based on different breakthrough time of the experiments). For the tracer experiment samples, NaCl concentration in each sample was measured with an electrical conductivity meter (Thermo Fisher). The metformin concentration in each sample was measured directly by a high-performance liquid chromatography with an ultraviolet detector (HPLC-UV, Agilent

Technologies) after filtering 1 mL samples through 0.22 μm filters(Briones & Sarmah, 2018b; Oertel et al., 2018). A Shimadzu HILIC column (250 × 4.6 mm, 5μm) was used with a liquid phase of acetonitrile and 10 mM ammonium acetate (pH adjusted to 3 by acetic acid). The volume ratio of acetonitrile and ammonium is 20:80. The flow rate of the mobile phase was 1.2 mL/min, and the injected sample volume was 20 μL; the UV detection wavelength was 233 nm. The retention time of metformin was 4.9 min. More experimental parameters and details are given in Table 4.3.

Table 4.3 Experimental conditions and parameters

#	Sand Type	Height (cm)	Diameter (cm)	Bulk density (g/cm^3)	Porosity	Directions	Pressure Head (cm)	Inflow rate (ml/min)	Initial Metformin Concentration (mg/L)
1	Fine	27	5	1.64	0.33	Top-down	3	1.600	10
2	Fine	27	5	1.59	0.31	Top-down	0	1.000	10
3	Medium	27	5	1.54	0.32	Top-down	3	16.800	10
4	Medium	27	5	1.63	0.29	Top-down	0	1.000	10
5	Mixed Variant 1	27	5	1.73	0.31	Top-down	3	1.700	10
6	Mixed Variant 1	27	5	1.76	0.26	Top-down	0	0.533	10
7	Mixed Variant 1	60	8	1.77	0.31	Top-down	3	6.087	10
8	Mixed Variant 2	27	5	1.74	0.30	Top-down	3	0.900	10
9	Mixed Variant 2	27	5	1.77	0.28	Top-down	0	0.406	10
10	Mixed Variant 2	60	8	1.74	0.30	Top-down	3	2.830	10
11	Coarse	24	2.4	1.55	0.30	Bottom-up	0	0.444	5
12	Coarse	24	2.4	1.62	0.27	Bottom-up	0	0.444	10
13	Coarse	24	2.4	1.54	0.32	Bottom-up	0	0.444	2.5
14	Coarse	24	2.4	1.55	0.32	Bottom-up	0	0.444	20
15	Coarse	24	2.4	1.56	0.33	Bottom-up	0	0.444	40
16	Coarse	24	2.4	1.67	0.28	Bottom-up	0	0.666	2.5
17	Coarse	24	2.4	1.60	0.29	Bottom-up	0	0.666	5
18	Coarse	24	2.4	1.60	0.27	Bottom-up	0	0.666	10
19	Coarse	24	2.4	1.67	0.26	Bottom-up	0	0.666	20
20	Coarse	24	2.4	1.58	0.29	Bottom-up	0	0.666	40
21	Coarse	24	2.4	1.66	0.26	Bottom-up	0	0.888	2.5
22	Coarse	24	2.4	1.63	0.26	Bottom-up	0	0.888	5
23	Coarse	24	2.4	1.59	0.29	Bottom-up	0	0.888	10
24	Coarse	24	2.4	1.55	0.30	Bottom-up	0	0.888	20
25	Coarse	24	2.4	1.58	0.29	Bottom-up	0	0.888	40
26	Coarse	24	2.4	1.59	0.29	Bottom-up	0	1.111	2.5
27	Coarse	24	2.4	1.53	0.31	Bottom-up	0	1.111	5
28	Coarse	24	2.4	1.61	0.29	Bottom-up	0	1.111	10
29	Coarse	24	2.4	1.66	0.26	Bottom-up	0	1.111	20
30	Coarse	24	2.4	1.58	0.30	Bottom-up	0	1.111	40

4.2.3 Metformin Transport Modelling and Data Augmentation

To estimate the parameters ($\widehat{\omega}$ in Figure 4.1) of metformin transport within the column system, the breakthrough curves of the transport experiments were modelled. The Hydrus-1D software package was chosen as the physics-based modelling tool for this study to fit the breakthrough curves of the tracer and metformin (Fan, 2022; Šimůnek & van Genuchten, 2008). This section presents the foundational assumptions and parameters crucial to the proposed framework.

The transport equation in the Hydrus-1D model can be given as follows:

$$\left(1+\frac{\mathrm{F}\rho\mathrm{K_d}}{\theta}\right)\frac{\partial \mathrm{C}}{\partial \mathrm{t}}+\frac{\rho}{\theta}\frac{\partial \mathrm{S^k}}{\partial \mathrm{t}}=\mathrm{D}\frac{\partial^2 \mathrm{C}}{\partial \mathrm{x}^2}-\mathrm{v}\frac{\partial \mathrm{C}}{\partial \mathrm{x}} \tag{4.1}$$

$$\mathrm{S^k}=(1-\mathrm{F})\mathrm{K_d}\mathrm{C_e} \tag{4.2}$$

The retardation factor R can be expressed as:

$$\mathrm{R}=1+\frac{F\rho K_d}{\theta} \tag{4.3}$$

Where F is the fraction of type-1 sorption; ρ is the bulk density of the porous medium $[ML^{-3}]$; K_d is the partition coefficient $[L^3M^{-1}]$; C is the solute concentrations in the liquid phase $[ML^{-3}]$; S^k is the type-2 site $[MM^{-1}]$; C_e is the solution concentration at equilibrium $[ML^{-3}]$; D is the dispersion coefficient $[L^2T^{-1}]$; v is the average pore-water velocity $[LT^{-1}]$; t and x are time $[T]$ and distance $[L]$, respectively; θ represents porosity. And $1/R$ is referred to as "relative velocity" in the main text (Šimůnek & van Genuchten, 2008).

In this study, the two-site conceptualization was used to describe the nonequilibrium sorption of metformin (Selim et al., 1977; van Genuchten & Wagenet, 1989). It is assumed

that the overall adsorption in the system can be categorized into two types: The first type ("Type-1") is instantaneous sorption, while the other type ("Type-2") indicates time-dependent kinetic sorption with a first-order reaction rate (Rao et al., 1979). The reciprocal of the retardation factor R can represent the relative travelling distance of a substance compared to water. Hence, it will be referred to as *relative velocity* below. Also, because relative velocity is one of the most intuitive metrics to evaluate the overall dynamics of metformin's transport in a 1-D system while retaining most of the crucial information, it is selected as the target variable in our study for both the causal effect estimation and neural network modelling. A 90% concentration of the initial value (C_0, ML^{-3}) marked the endpoint for porous medium transport.

Data augmentation refers to the significant enhancement of data volume and diversity available for training data-driven models without the need for new data collection. This process can mitigate overfitting and bolster model performance and robustness (van Dyk & Meng, 2001). Under the context of investigating metformin transport in sandy media, it is essential to preserve the characteristics of these processes within the augmented data. With this consideration, we employed a grid-traverse method in tandem with certain constraints to produce parameter combinations for data augmentation, ensuring the exclusion of improbable parameter sets, considering a series of constraints to ensure the augmented dataset faithfully represents metformin transport in a sandy medium, as opposed to the arbitrary transport of a substance in a random medium. As a result, we assembled an augmented dataset with 54,869 parameter combinations, which were subsequently normalized using 0-1 normalization. This normalization aids neural network training and generates normalized causal effect metrics, enabling comparison and easier

interpretation. To prevent data leakage and maintain the integrity of our findings, this dataset was then divided into two subsets for subsequent causal inference and neural network training: 30% for causal inference, containing 16,460 instances, and 70% for neural network training, containing 38,409 instances. The batch runs of the Hydrus-1D model were performed using Phydrus, a Python implementation of Hydrus-1D (Collenteur et al., 2020).

Based on fitting 60 breakthrough curves (including 30 sets for tracer and 30 sets for metformin), a reasonable range of all parameters can be deduced from the results. Then, based on the ranges, a total of 200,000 instances were generated through Phydus. Any instances that failed to converge, implying the combination of conditions and parameters are unlikely reasonable, are removed.

The particle density (ρ_s) of a porous medium can be calculated using the bulk density (ρ_b) and the porosity (θ) of the medium. The equation is: $\rho_s = \rho_b/(1-\theta)$. The data is then filtered based on the particle density range (Bear, 2013), which should be between 2.2 and 2.6. Following this, range limits for hydraulic conductivity, adsorption coefficient, porosity and bulk density were set to screen out outliers in the dataset. The final constraints used during data augmentation include 1) typical ranges for sandy porous media properties such as porosity and bulk density; 2) metformin transport parameter distributions acquired from the column experiment; and 3) ensuring the simulated processes can converge within the model framework. The final dataset is available in our data repository.

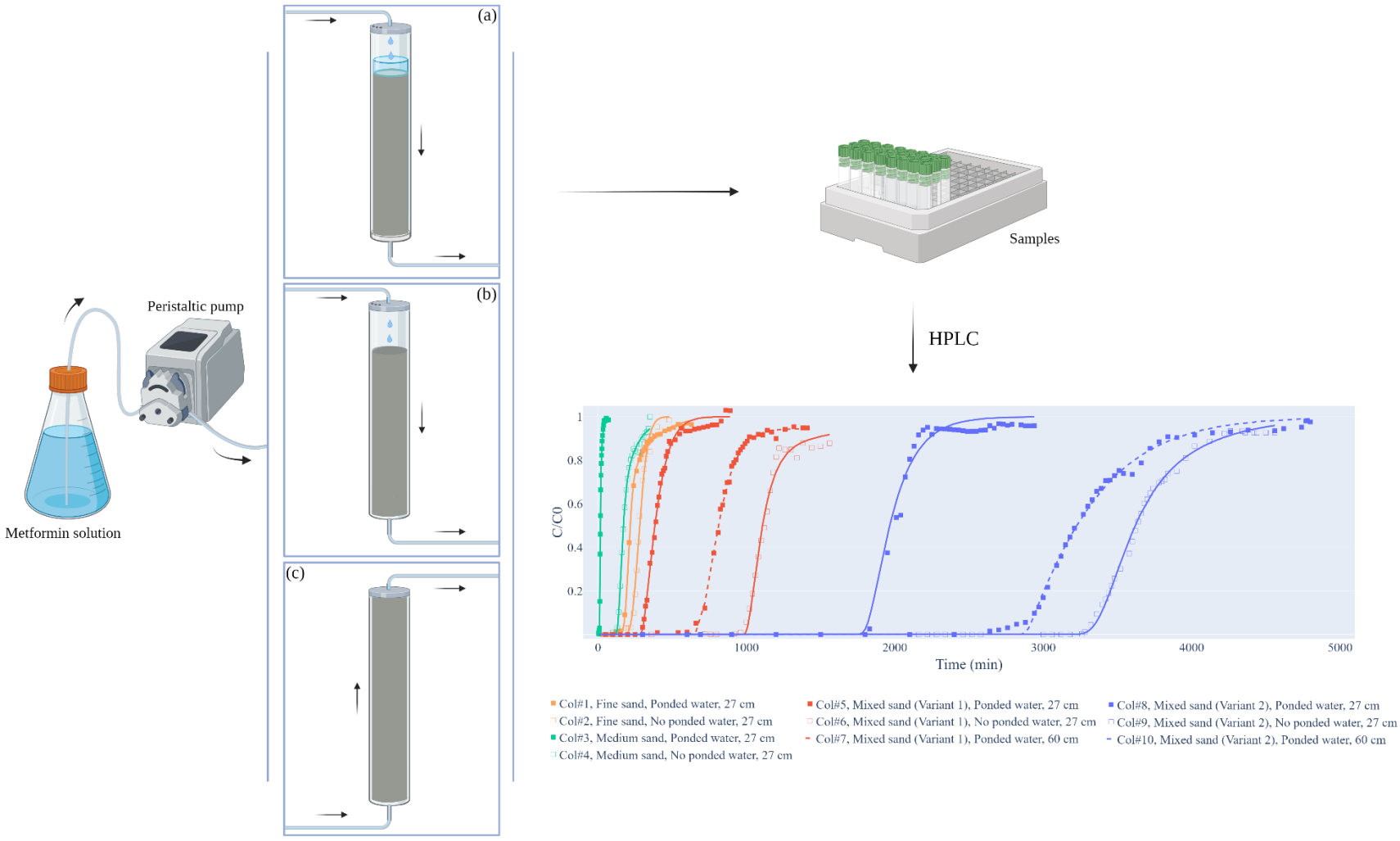


Figure 4.2 Schematic diagram of the metformin column transport experiment. Arrows indicate the flow direction. Setting (a): Saturated top-down flow with 3 cm ponded water; Setting (b): Unsaturated top-down flow with no ponded water; Setting (c): Saturated bottom-up flow; (d) Metformin breakthrough curves under top-down saturated and unsaturated flow conditions. Solid and open squares represent experimental data with open shapes correspond to experiments without ponded water; solid and dashed lines represent fitted breakthrough curves with dashed lines represent experiments in 60 cm columns.

4.2.4 Causal Inference

The first step of SCM is to construct a DAG for causally relevant variables. Based on the domain knowledge, 14 variables including the porous medium's physical, hydraulic, and transport parameters, along with experimental boundary conditions, were included in the causal diagram. Specifically, they are:

- θ: Porosity;
- PD: Particle Density;
- S: Saturation;
- L: Travel distance;
- C: Concentration;
- K_d: Partitioning coefficient;
- K_s: Hydraulic conductivity;
- F: Type-1 sorption fraction;
- D: Dispersivity;
- α: First-order reaction rate for kinetics sorption;
- H: Ponded water depth;
- q: Water flow flux;
- g: Travel direction;
- $1/R$: Metformin velocity relative to water.

Given the causal relationships depicted in the causal diagram, suitable sufficient sets—specific variables that encapsulate the causal effect—can be identified using graph-based procedures, primarily the backdoor adjustment.

In our study, two different estimators to capture both linear and non-linear causal effects were used: a classic linear model and a machine learning-based causal estimator CausalForestDML, previously named DMLOrthoForest (Chernozhukov et al., 2017; Foster & Chilton, 2003). Approximately one-third of the augmented dataset (~18,500 instances) was used here for causal regression analysis. An additive model to represent the dynamics between different parameters and metformin's relative velocity ($\frac{1}{R}$) in porous media was realized as the causal prior for providing causal information into the downstream neural network model.

SCM framework and three refutation tests in the study are available within the DoWhy package, an open-source Python package for causal inference (Sharma & Kiciman, 2020). An open-source library, EconML, provides an implementation of the machine learning estimator CausalForestDML (Battocchi et al., 2019). As a benchmark, the estimators have been applied to multiple public datasets with known ground truth in previous studies.

4.2.5 Causal Prior Embedding

To embed the extracted causal prior into the neural network, two methods were used in this study. The first method of causal prior embedding is causal regularization (Kancheti et al., 2022). On top of the ordinary regularization techniques, which aim to reduce overfitting, causal regularization introduces an additional penalty term into the learning process to encourage the model to align its learning with our prior understanding of the causal structure of the problem. The penalty term is determined by the L1 norm of the discrepancy between 1) the Jacobian matrix of the model with respect to its input and 2) the derivative

of the causal prior. In a linear context, the derivative of the causal prior corresponds to the Average Treatment Effect (ATE) for each influential feature. The regularization term was applied in such a way that the network's learning process would strive to minimize the difference between its own understanding of the system and the known causal relationships. This step can improve the model's interpretability without significantly damaging model performance, by making it adhere to known causal structures (Kancheti et al., 2022; Suryadi et al., 2023; Chattopadhyay et al., 2019).

Causal regularization imposes a constraint on the learning process of the neural network, nudging it towards better alignment with prior causal knowledge. This process involves matching the gradients of the NN's causal effects with the gradients of a domain prior function. The regularization term can be given as:

$$R = \frac{1}{N}\sum_{i=1}^{N} \max\left(\left|\left|A_i - B_i\right|\right|_1 - \epsilon, 0\right) \tag{4.4}$$

Here, N represents the number of inputs; A_i and B_i are the i-th rows of matrices A and B respectively; Matrix A is the *Jacobian* of the network's output with respect to its input; Matrix B represents a matrix of derivatives that stand for the known causal relationships; the error term ϵ is a small constant introduced to allow flexibility in the matching between A and B.

$\left|\left|A_i - B_i\right|\right|_1$ denotes the $L1$ norm, which quantifies the absolute difference between the corresponding elements of matrices A and B. In essence, the regularizer encourages the

gradients of the network's output with respect to its input to be close to the known causal effects, guiding the model towards physically meaningful solutions.

In addition to causal regularization, a novel approach was developed to incorporate causal knowledge into the learning process of neural networks, which was refer to as causal weight initialization. Neural networks, at their core, are composed of layers of nodes (or "neurons") that are interconnected through "weights." These weights are essentially the parameters that the network adjusts during the learning process to improve its predictions. The initialization of these weights – the values they are given before the learning process begins – can significantly influence the network's learning trajectory and final performance. In the case of causal weight initialization, the prior knowledge about the causal relationships between different factors to guide this initial assignment of weights was used (Kassani et al., 2022; Luo et al., 2020). Specifically, the initial weights in the input layer of the network are set to reflect the ATEs, a measure of the causal effect of each input feature.

The causal weight initialization procedure leverages the Average Treatment Effect (ATE) values as priors in defining the initial weights of the neural network. In this study, for each input feature, the initial weight values are generated by drawing from a Gaussian distribution, with the mean equal to the ATE value for that input feature and a small standard deviation, signifying the uncertainty around the ATE. This process creates an array of weights with dimensions equal to the number of inputs by the number of nodes in the layer and hence can be used for input layer weight initialization. This way, the learning process starts from a state that's consistent with the known causal effects, offering a

promising direction to enhance the interpretability and reliability of model predictions. The algorithm for ATE calculation is based on a Taylor series expansion of the neural network output and employs both first-order and second-order gradients in its computations. The ACE is essentially the difference between two CATEs, as shown in the following pseudo-code:

Algorithm: Calculation of Average Causal Effect (ACE) learned by the neural network

Result: Expected value of Y given t (E[Y|t]) for each t = α
Inputs:
- Function f that takes t as an argument
- The range of t: from 'low' to 'high'
- Number of interventions: n
- Mean of the data: μ
- Covariance matrix of the data: cov
Procedure:
1. Initialize α = low, a list IE = []
2. While α is less than high:
  2.1 Set the i-th element of μ to α
  2.2 Calculate the function f at μ, append 1/2 of this value to the list IE
  2.3 Add the trace of the product of matrix multiplication (second derivative of f at μ, cov) to the last element of IE
  2.4 Increment α by (high - low) / n
3. Return the list IE

A Neural Architecture Search (NAS) parameterization experiment was conducted to assess the impact of various hyperparameters on model performance. A range for each parameter was predefined and models were trained with all combinations of these parameter values. Four standard neural network hyperparameters were considered: learning rate, number of layers, number of nodes, and the choice of activation functions. In addition to these, two causal embedding parameters were also taken into account: 1) the causal regularization parameter, represented as λ, and 2) a Boolean parameter indicating whether the model's input layer weights were causally initialized. The models were evaluated not only on standard performance metrics such as convergence speed and performance on a test set but also on how well the model retained causal information derived from the prior. This validation set was separate from the training and testing sets to avoid data leakage and ensure rigorous evaluation. All models were trained and evaluated on the same training and testing sets, providing a consistent basis for comparison across different hyperparameter combinations. The neural network model was developed through the open-source deep learning framework PyTorch 1.13.1, and the NAS parameterization was conducted through an open-source AutoML framework Neural Network Intelligence (NNI) 2.10 (Microsoft, 2021).

4.3 Results and discussion

4.3.1 Metformin Transport in Sandy Media and Causal Interactions Within

The breakthrough curves of metformin are given in Figure 4.2(d) for settings (a, b) and Figure 4.3 for settings (c). Observed experimental data are denoted by squares and circles, while the fitted curves are in solid or dashed lines. The modelled curves showed good alignment with the original experimental data, with their R^2 ranging from 0.970 to 0.999, indicating that the selected physics-based model's capability describes metformin's sorption in sandy columns. Meanwhile, as shown in Figure 4.2(d), it is noticeable that metformin's residence time increased in the following order: medium sand (Col# 3,4), fine sand (Col# 1, 2) and mixed sand (Col# 5-10). The reason that metformin was stranded for the longest time within the mixed sand columns is due to the heterogeneous nature of the mixed sand, which increased the transport system's complexity and impeded the transport of metformin, in contrast to more homogeneous media like medium and fine sands (Beven, 1996; Simmons et al., 2001; Gelhar & Axness, 1983). Stranding time also varies between different mixed sand variants. In columns packed with mixed sand variant 1 (Col# 5-7), the breakthrough time for metformin ranged from 510 to 1,450 min. In columns with mixed sand variant 2 (Col# 8-10), the breakthrough time ranged from 2,230 to 4,200 min. Such a difference was due to the higher proportion of finer particles in the latter variant, as shown in Table 4.1. Moreover, for the columns filled with the same sandy medium, the breakthrough time of metformin was longer under unsaturated condition than under saturated condition. On the other hand, metformin's breakthrough time under bottom-up saturated flow conditions in coarse sands (Col #11–30, ranging from 72 to 295 minutes) was more dependent on the flow flux, as shown in Figure 4.3.

Table 4.4. Estimated parameters of the sandy columns

#	K_s (cm/min)	D (cm)	F	K_d (L/kg)	α	$1/R$
1	0.075	0.156	0.646	0.231	0.0081	0.568
2	0.066	0.159	0.836	0.161	0.0148	0.648
3	0.670	0.806	0.221	0.076	0.1104	0.962
4	0.126	0.468	0.483	0.264	0.0115	0.798
5	0.085	0.178	0.773	0.555	0.0237	0.345
6	0.028	0.511	0.854	0.105	0.0023	0.238
7	0.100	1.198	0.769	0.415	0.0004	0.223
8	0.049	0.390	0.655	0.370	0.0857	0.151
9	0.027	0.209	0.297	1.130	0.0060	0.101
10	0.057	0.608	0.835	0.606	0.0030	0.190
11	1.025	0.174	0.119	0.085	0.0139	0.722
12	0.906	0.121	0.179	0.070	0.0187	0.728
13	1.107	0.094	0.175	0.085	0.0148	0.883
14	1.181	0.104	0.216	0.070	0.0210	0.916
15	1.010	0.095	0.162	0.036	0.0218	0.943
16	0.820	0.212	0.145	0.074	0.0185	0.881
17	1.391	0.102	0.164	0.072	0.0282	0.910
18	0.931	0.150	0.180	0.085	0.0273	0.904
19	1.020	0.438	0.242	0.074	0.0268	0.877
20	0.977	0.217	0.083	0.019	0.0088	0.990
21	1.012	0.358	0.111	0.058	0.0261	0.893
22	1.116	0.199	0.199	0.100	0.0281	0.882
23	0.980	0.118	0.157	0.081	0.0352	0.930
24	1.025	0.154	0.186	0.075	0.0336	0.928
25	1.135	0.157	0.177	0.052	0.0387	0.945
26	1.174	0.163	0.126	0.075	0.0309	0.926
27	1.053	0.142	0.089	0.089	0.0379	0.962
28	0.968	0.289	0.152	0.064	0.0317	0.931
29	0.908	0.408	0.103	0.054	0.0383	0.930
30	1.008	0.161	0.177	0.046	0.0413	0.941

* K_s : Hydraulic conductivity, D: Dispersivity; F: Type-1 sorption fraction; K_d: Partitioning coefficient; α: First-order reaction rate coefficient for kinetics sorption; $1/R$: Metformin velocity relative to water.

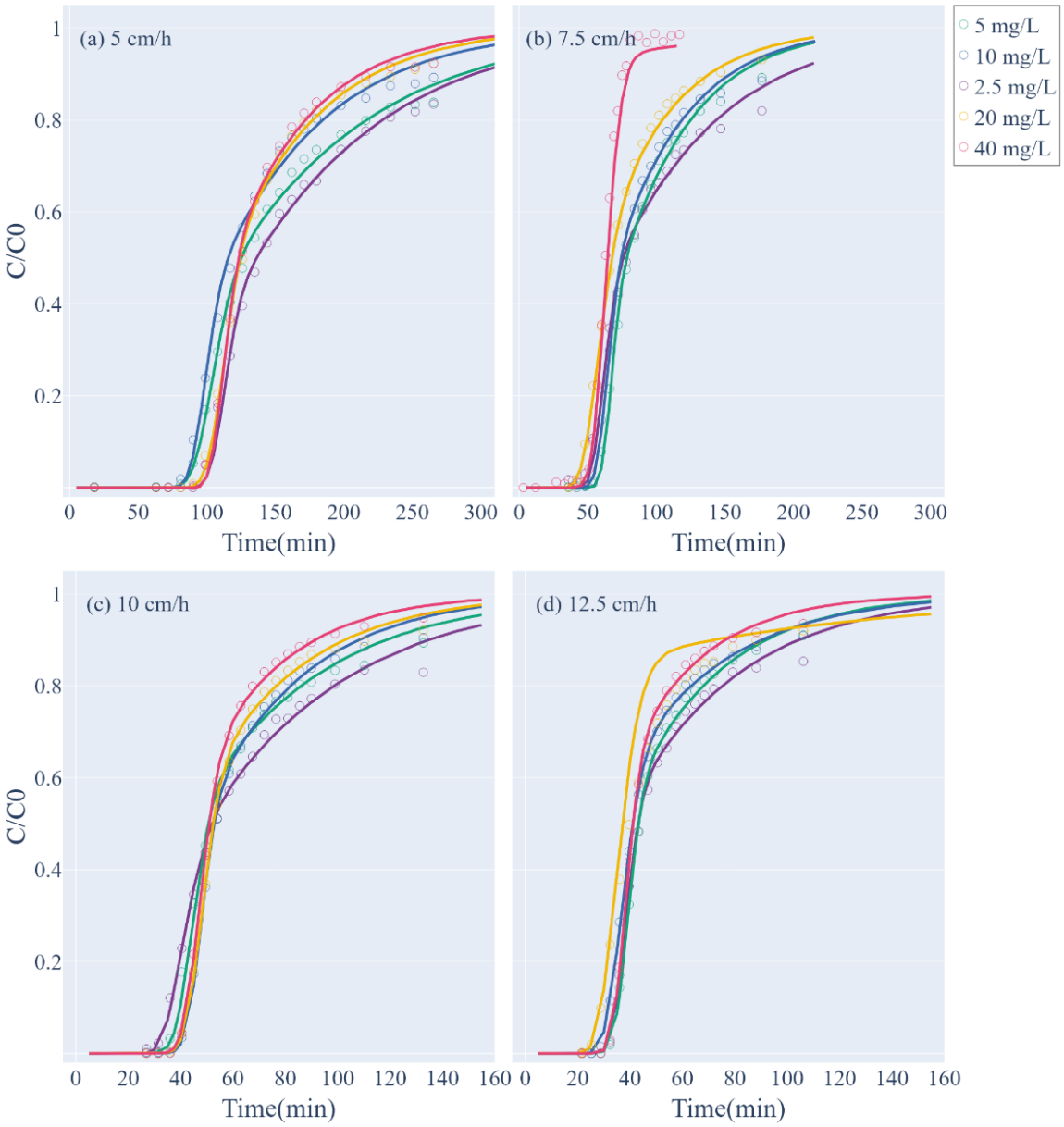


Figure 4.3 Metformin breakthrough curves under bottom-up flow conditions with different concentrations.

Based on the experimental data, the fitted parameters from the Hydrus-1D model in individual columns are given in Table 4.4 and were incorporated in the original experimental dataset. Variability across the transport parameters such as type-1 sorption fraction (F), hydraulic conductivity (K_s) and dispersivity (D) suggests that the original experiment settings covered various scenarios. Hence, the new dataset, which includes both the experimental and model-fitted-parameters, was deemed "seed data", and was subsequently used to generate new synthetic data for the causal analysis and deep learning model training, as previously described.

Based on the causal inference dataset as described in Section 4.2.3, a weighted causal diagram was constructed to showcase the results of causal effect estimations, as in Figure 4.4. This causal graph seeks to demonstrate the complex interplay of various factors that contribute to solute transport in a porous medium. Some factors and causal interactions are discussion worthy. The first factor of interest is the type-1 Sorption Fraction. The fitted F in the original seed data varied from 0.08 to 0.91, with an observable higher value in finer sands. Type-1 sorption fraction is a variable influenced by many other factors, such as porosity, particle density and saturation status, instead of merely a dependent variable of sand type. Specifically, porosity emerges as a dominant factor, with a normalized causal effect of 0.0242. This corroborates the concept that increased porosity tends to provide more sorption sites for solute interaction, affecting the partitioning of the solute between the solid and liquid phases (Rao et al., 1979; van Genuchten, 1980). The saturation status and particle density have smaller effects on type-1 sorption fraction (0.0082 and –0.009). The underlying mechanism of the minor shift brought by saturation status could be that the water-filled pores might slightly facilitate greater interaction between the contaminant and

the solid phase, leading to a moderately higher instantaneous sorption fraction (Bear & Cheng, 2010). On the other hand, the minor negative effect from particle density corresponds to that when all the other factors are held constant, a medium with higher particle density implies a slightly limited availability of sorption sites, hence might slightly reduce the fraction (van Genuchten & Wagenet, 1989; Freeze & Cherry, 1979).

Another pair of the treatment-outcome variables of interest is dispersivity D and travel distance L. Their relationship has been a topic of active discussion since the mid-twentieth century when researchers first documented the interaction between these two parameters. Two interpretations are commonly held in the field: one believes heterogeneity causes such a scale effect (G. Gao et al., 2010; Gelhar et al., 1992), and the other suggests that the scale effect might be a technical artifact (Domenico & Robbins, 1984). Although no consensus has been reached, it has been widely accepted that dispersivity is a distance-dependent parameter with a clear physical meaning (Bromly et al., 2007; Schulze-Makuch, 2005). Recently, to interpret the relationships between dispersivity and distance, researchers investigated multiple experimental datasets and found that dispersivity is still a good descriptor of a transport system (Younes et al., 2020; You & Zhan, 2013). In our case study, distance indeed shows its influence on dispersivity. Albeit a lesser influence compared to other factors, with a normalized weight of 0.023, it signifies a 2.3% shift in dispersivity when the transport distance ranges from its minimum observed value (20cm) to its maximum (160cm). The impact of distance on dispersivity stems from the nature of solute transport over different scales. As the travel distance increases, the solute particles have a higher likelihood of experiencing diverse flow paths and velocity variations within the system, leading to more extensive spread and, hence, greater dispersivity. This highlights

the spatial aspect of solute transport, underscoring that even when other conditions remain constant, an increase in the distance traversed by the solute can lead to a small yet observable increase in dispersivity (Menzie & Dutta, 1989; Schulze-Makuch, 2005). Thus, both porosity and distance contribute in their unique ways to the overall dispersivity in the system, although their impacts may vary in magnitude. The "scale effect" from the travel distance to the relative velocity was also quantified (-3.51%), indicating a measurable impact of the scale of the system on the transport process of metformin. It is also consistent with numerous empirical observations and theoretical models suggesting a slowdown in contaminant transport as the transport distance increases (Domenico & Robbins, 1984).

Among all the factors affecting the target variable, the saturation status (S) stands out as a dominant factor, contributing a positive impact of 0.1625, in which a system shift from unsaturated to saturated conditions increases the relative velocity remarkably. This estimation underscores the role of saturation status in enhancing the transport speed of metformin, possibly by increasing the continuity of the water phase and thus facilitating the transport process (Freeze & Cherry, 1979). Another noteworthy influence comes from the adsorption coefficient K_d, which exerts a substantial negative effect of -106.44%. This suggests that an increase in the adsorption coefficient, which reflects the tendency of metformin molecules to adhere to the sand particles, considerably decelerates the relative velocity, implying an intensified retardation effect on the transport (Šimůnek & van Genuchten, 2008). The first-order reaction rate coefficient for type-2 sorption also has a -0.1447 normalized causal effect on the relative velocity. It implies a faster rate of this sorption reaction would result in a greater proportion of the metformin adhering to surfaces and not moving freely in the water, resulting in a reduction in its overall transport velocity

(Maraqa et al., 2011). More detailed causal estimation and refutation results are given in Tables B.1 and B.2 in the Appendix B, including the backdoor adjustment sets for each causal link.

Transport parameters in different types of sandy columns had distinctive distributions. Hydraulic conductivity (K_s) in the coarse sand columns (Col #11–30) was noticeably higher than in other sandy columns (0.82–1.18 cm/min), with their dispersivity (D) comparably lower, probably due to their slightly shorter column length (24 cm). The average relative velocity in the first ten column experiments was 0.42, nearly half of it in the coarse sandy columns #11–30 (0.90). On the other hand, the partitioning coefficient in Experiment #11–30 (coarse sand columns) was also noticeably lower, indicating there was not as much sorption occurring in those experiments as in Experiment #1–10.

Particle density (PD in the Figure 3) has been identified as a parent node influencing hydraulic conductivity, dispersivity, adsorption coefficient, and the fraction of Type-1 sorption. When all the other factors are held constant, a medium with higher particle density can reduce the pore space available for fluid flow, affecting hydraulic conductivity and dispersivity, and it can influence the availability of adsorption sites, thus affecting the adsorption coefficient and Type-1 sorption fraction (van Genuchten & Wagenet, 1989; Freeze & Cherry, 1979). That prior knowledge was captured in those noticeable weighted causal effects from particle density to adsorption coefficient K_d (0.334) and hydraulic conductivity K_s (–0.673). On the other hand, porosity (θ), which signifies the void spaces within the medium, also influences hydraulic conductivity, dispersivity, the adsorption coefficient, and Type-1 sorption fraction. Higher porosity generally affects hydraulic conductivity and dispersivity due to more interconnected void spaces. Similarly, a higher

porosity could increase the number of available sites for sorption, influencing both the adsorption coefficient and Type-1 sorption fraction(Bear, 2013; van Genuchten & Wagenet, 1989; Freeze & Cherry, 1979). Hence, such insights were also reflected in the corresponding weights in Figure 3 (-0.071 and 0.209 for K_d and K_s, respectively).

For dispersivity, the saturation status (S) appears to have the most substantial impact, with a normalized weight of 0.0483. As a binary variable, a change in the degree of saturation from unsaturated (0) to saturated (1) results in a 4.83% change in the dispersivity value within its observed range. This aligns with the fact that the saturation state of the system can drastically influence the transport dynamics of the solute, with a saturated system typically having a greater dispersivity. Particle density also exerts a noticeable influence on dispersivity, with an estimated causal effect of 0.0371. This coincides with the understanding that denser particles reduce pore spaces and affect the diffusion and flow of solutes, impacting their spread (Bear, 2013; Freeze & Cherry, 1979). Porosity, another key factor, presents a smaller effect of 0.0127, which corresponds with the fact that an increase in porosity has the potential to increase the dispersivity due to that higher porosity typically provides a greater volume for fluid flow and more complex flow pathways, which can lead to enhanced mixing and dispersion of solutes (Schulze-Makuch, 2005; Menzie & Dutta, 1989).

Hydraulic conductivity K_s presents minor effects of 3.39%, suggesting the effect of the connectivity of the pore spaces on relative velocity. Additionally, the negative influences from the type-1 sorption fraction (–4.23%), flux (–2.54%), and distance (–3.51%) suggest their roles in impeding metformin transport. The type-1 sorption fraction, indicating the fraction of metformin that adheres to the soil particles instantaneously, and flux, reflecting

the flow rate of water, logically affect the transport speed by increasing the retention of metformin and reducing the driving force, respectively. The negative effect of flux conforms to the expectation that the relative velocity of the solute will be slower when all the other conditions remain the same. It is because that in a higher flux situation, water molecules move more quickly through the soil and metformin molecules may not travel as quickly, which reflected in a lower relative velocity. As the distance that metformin molecules travel increases, various influencing factors like diffusion, adsorption, and degradation can come into play, reducing the speed at which they move through the system.

An *F* range from 0.38 to 0.45 for oxytetracycline in sandy loamy soil mixed with polyamide was reported by Li.(J. Li et al., 2021). Zhou et al.(Zhou et al., 2019) reported that no type-1 sorption occurred during the transport process of ciprofloxacin in hematite-coated-sand-packed columns. In their research on the transport of aniline and nitrobenzene, Zakari et al. (Zakari et al., 2019) suggested that type-1 sorption of the organic compounds occurs on the sediment particle surface while type-2 sorption occurs on the micropore surface. They also found that the *F* value dropped significantly when the velocity surpassed a certain threshold in another study on the transport of bisphenol-A in sandy columns(Zakari et al., 2016). The authors explained the phenomenon by suggesting that more BPA on the type-1 site (instantaneous) was “driven out” by fast water flow. On the other hand, during an investigation on the transport of seven phthalates in biosolid-amended soil, Sayyad et al. reported a low *F* value range near zero and emphasis the parameter’s relationships with molecular size, substitution pattern of molecules, length of the carbon chain and Log K_{ow} (Sayyad et al., 2017), which can meet the research conducted by Brusseau et al. and Maraqa et al.(Brusseau et al., 1991; Maraqa et al., 2011). Some

studies also reported correlations between pH and F. Xing et al.(Xing et al., 2020) suggested the instantaneous sorption of tetracycline increased with the presence of colloids, implying that colloids might shorten the migration time of tetracycline in sandy media.

The relative velocity of metformin in the sandy columns was within the same range shared by some other PPCPs such as bisphenol-A, tetracycline, and thiacloprid (Wei et al., 2021; Rodríguez-Liébana et al., 2018; Zakari et al., 2016) and was on the higher end of the spectrum (a 0.74 average across all the experiments). Given metformin's insensitivity to photolysis and hydrolysis and the degradation rate of metformin is contingent upon specific environmental factors (Caldwell et al., 2019), metformin may exhibit a longer persistence in the groundwater compartment and may still cause a considerable travel distance, even though the velocity of groundwater is generally slow compared with surface water flows. The information above suggested a long-range transport potential in the groundwater (Griebler & Lueders, 2009), which partially explained the ubiquitous occurrence of metformin in natural water bodies.

4.3.2 Embedding Causal Prior into a Neural Network

Figures 4.4 (a) present the results of all 1,440 neural networks parameter searching runs, detailing overall, top 10%, and bottom 10% outcomes, respectively. Certain experiments exhibited remarkable performance, with R-squared values peaking at 0.98 and root mean square error (RMSE) reaching lows of 0.02 on the test set, affirming the problem's learnability while positioning the dataset as an ideal platform with a balanced complexity, for the exploration of causal embedding techniques. To facilitate the upcoming discussion related to activation functions, a visualization of the curves for the five activation functions investigated in this study is given in Figure 2.1 in the Literature Review section along with their corresponding equations and detailed exploration of their respective advantages and disadvantages.

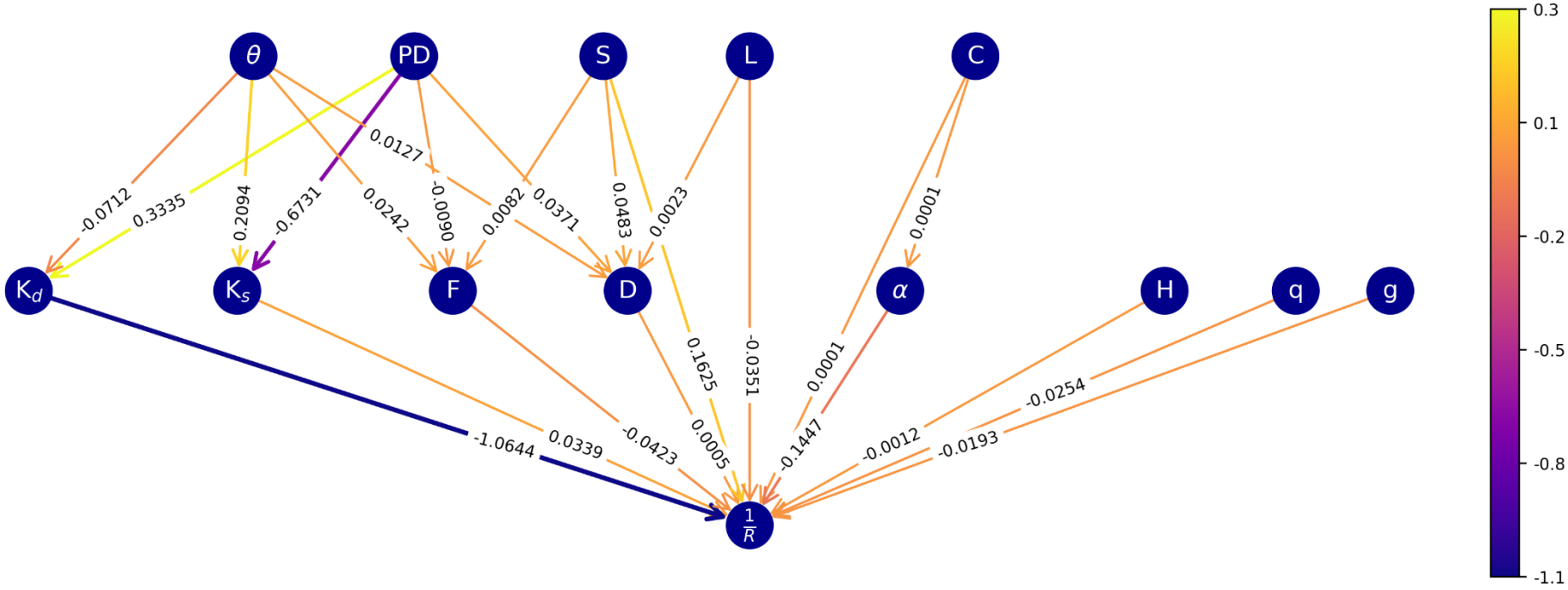


Figure 4.4 A DAG with the weighted edge indicates the estimated causal effect of different causal links. θ: Porosity; *PD*: Particle Density; *S*: Saturation; *L*: Travel distance; *C*: Concentration; K_d: Partitioning coefficient; K_s: Hydraulic conductivity; *F*: Type-1 sorption fraction; *D*: Dispersivity; α: First-order reaction rate for kinetics sorption; *H*: Ponded water depth; *q*: Water flow flux; *g*: Travel direction; *1/R*: Metformin velocity relative to water. The weights associated with each edge correspond to its normalized estimated causal effect and are comparable across all the interactions.

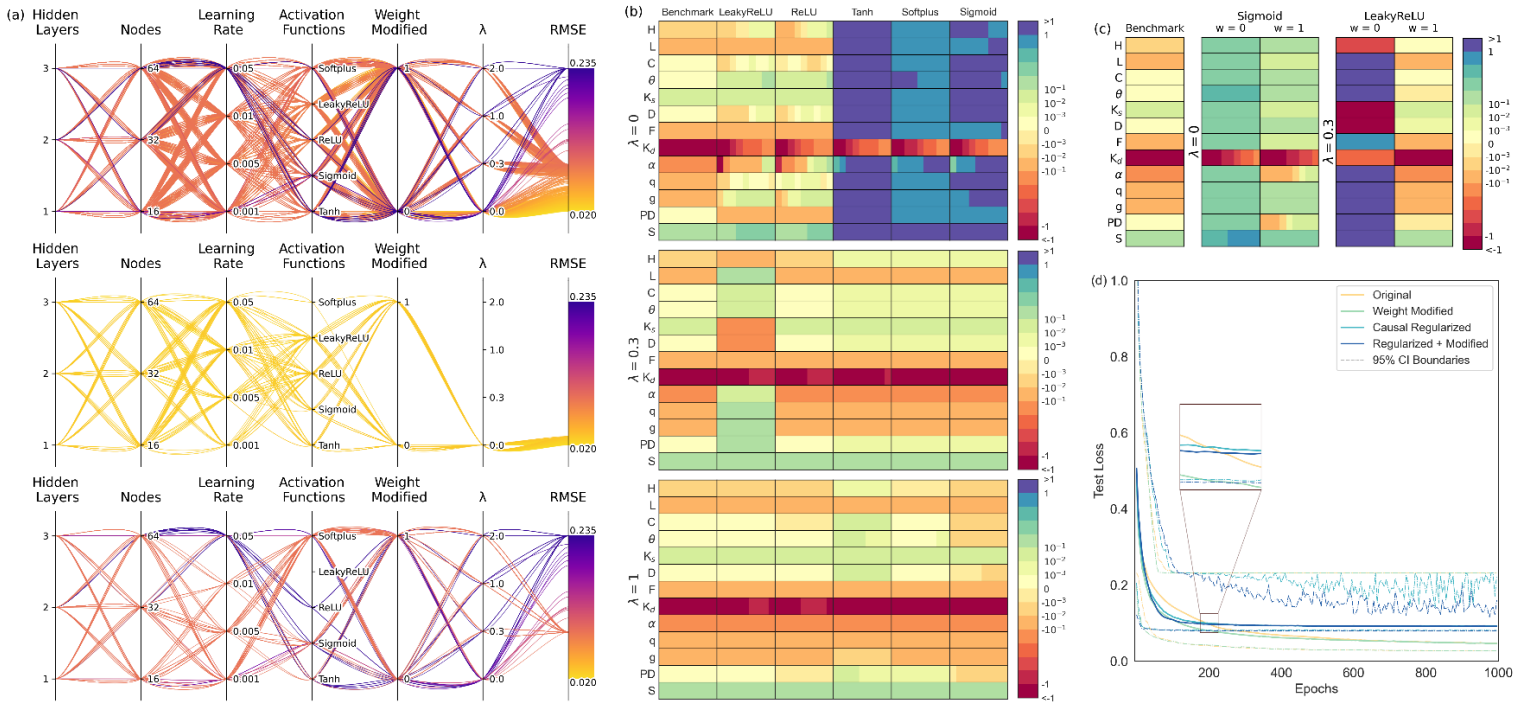


Figure 4.5 Overview of experimental results and causal prior retention. (a) RMSE results from all 1,440 experiments; top 10% performing experiments and bottom 10% experiments; (b) Causal retention heatmap for experiments with causal regularization, each cell represents the Conditional Average Treatment Effects (CATEs) at evenly spaced intervals of the normalized input variables; (c) Two particular cases of causal retention, visualized in causal retention heatmaps. On the left: two-layer network with 64-node hidden layers, Sigmoid, low learning rate (0.001), without causal regularization; On the right: a three-layer network 64-node hidden layers, LeakyReLU, high learning rate (0.05), with a minor causal regularization strength (λ=0.3); (d)Test loss curves for models with different causal embedding techniques.

The optimal hyperparameter combination relies on a subtle interplay between network structure, hyperparameters, and the choice of activation function. Most notably, networks comprising two or three layers, each with more than 16 nodes, coupled with LeakyReLU or ReLU activation functions, frequently yielded top-performing models, dominating the top 1% and 10% of experiments. LeakyReLU, in particular, proved to be a strong facilitator for networks of increased complexity, underlining its robustness as one of the most popular activation functions at the moment (B. Xu et al., 2015). On the other hand, some configurations resulted in less satisfying performance. Models employing the Softplus activation function frequently ranked in the bottom 10% of our experiments (90 out of 144). This result is consistent with known characteristics of the Softplus activation function. Specifically, it is prone to a phenomenon known as saturation, where it outputs values close to zero for small inputs. This behaviour can lead to near-zero gradients, which can subsequently hinder the performance of the model (B. Xu et al., 2015).

The interplay between various activation functions and causal embedding techniques revealed intriguing phenomena. As depicted in Figure 4.5(b), the benchmark column represents the originally embedded causal prior, with individual cells in the heatmap showcasing the Conditional Average Treatment Effects (CATEs) across evenly spaced intervals of the normalized input variables. Each cell value was computed by averaging across all experimental conditions involving different activation functions. This study utilized a logarithmically scaled colormap to scrutinize causal prior retention, particularly

interested in the effectiveness of both techniques when dealing with a simple additive causal prior with a mix of large and small terms— a scenario frequently encountered in environmental engineering and science disciplines. From the visualization, it can be observed that, in general, the causal regularization method can firmly introduce causal information into the model most of the time. However, there were still many signs of instability. For instance, LeakyReLU, despite its capability to capture a significant portion of the causal information without modification, exhibited the least retention of causal information among all activation functions when a minor causal regularization strength (λ=0.3) was applied. Interestingly, this deficiency in causal retention did not affect the overall performance and stability of the network, and it was not observed in ReLU, another piecewise linear activation function. One assumption is the discrepancy in behaviours under minor regularizing strength may be linked to LeakyReLU's non-zero gradient for negative inputs, unlike ReLU's zero gradient. This structural difference complicates LeakyReLU's task of matching the constant rate of change in output with respect to input (as enforced by the regularization term) when the regularizing strength is insufficient, leading to poorer causal information retention. When regularization strength increased to an intermediate level (λ=1), LeakyReLU matched the performance of other functions, overcoming its initial causal retention deficit. This improvement can be attributed to LeakyReLU's ability to handle complex transformations due to non-zero outputs for negative inputs and its resistance to the "dead neuron" problem, keeping the weight-

updating process active even under stronger regularization. On the other hand, some of the sacrifices on both performance and stability can also be observed in Figure 4.5(d), which depicts different model training processes by plotting the model's metric (RMSE) on the test set over epochs. The lower the metric, the better the model converged. The test losses of models utilizing causal regularization methods (navy blue and cyan curves) were restrained, as evidenced by their flattened curves higher than those without causal regularization (orange and green curves) after 200 epochs. These causal regularized experiment curves also exhibited fluctuating upper bounds, corresponding to the performance at the 95th percentile in each epoch. This fluctuation indicates the model's struggle to balance minimizing the MSE loss with matching the derivative of the causal prior, which is an inherent challenge of the regularization techniques. In contrast, such observations were not found for experiments without causal regularization, suggesting that, generally, those causal regularized experiments faced more difficulties in achieving convergence.

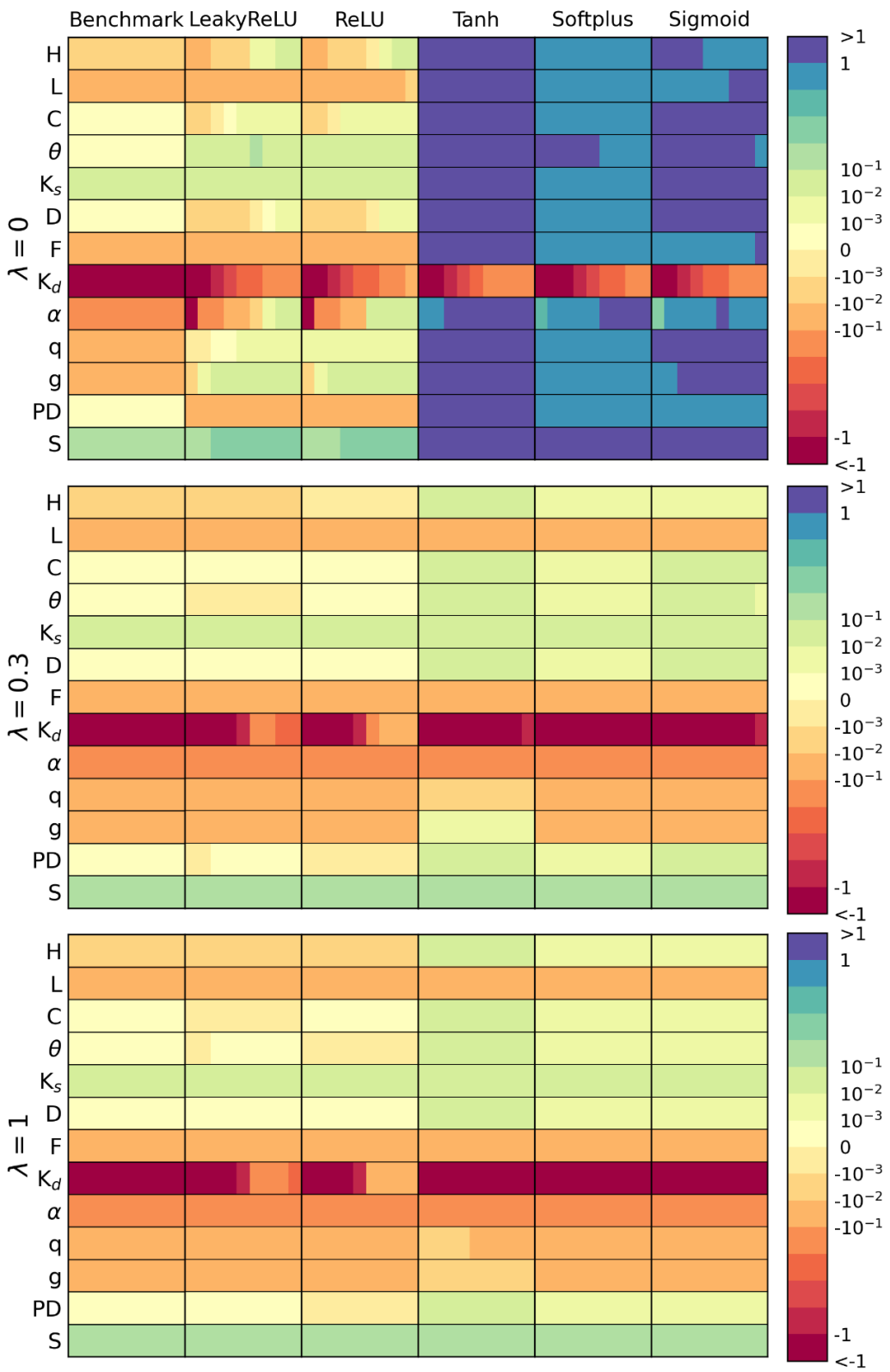


Figure 4.6 Causal retention heatmap for experiments with causal weight initialization.

Meanwhile, the causal embedding technique proposed and developed in this chapter, termed “causal weight initialization,” appears to influence the models’ learning processes in a distinct way. In contrast to causal regularization, which imposes an additional constraint on the learning process by encouraging the model’s derivative to align with the derivative of the causal prior, causal weight initialization serves as a form of guided starting point. It sets the initial model weights to reflect the known causal relationships, thereby orienting the learning process towards these causal patterns from the very beginning. This can potentially lead to more efficient convergence and causal retention, even without the explicit constraint imposed by causal regularization. As shown in the left panel of Figure 4.5(c), causal weight initialization proves effective for weight exploration in a Sigmoid-connected network, revealing some causal characteristics, even without explicit causal regularization. Interestingly, as depicted in the right panel of the same figure, a LeakyReLU-connected network utilizing causal regularization (λ=0.3) initially struggled to retain causal information, but the introduction of causal weight initialization enabled the network to gain causal information. This pattern was also noted with Softplus and Tanh. This can be attributed to the initial bias introduced by causal weight initialization. Specifically, this technique provides the network with a ‘head start’ towards the known causal relationships. In high variance scenarios brought on by flexible activation functions such as LeakyReLU, Tanh, and Softplus, such bias forms part of the bias-variance trade-off, and, overall, proves beneficial for the training process. Even if the neural network

models struggle with retaining causal information, causal weight initialization can guide the learning processes towards representing the designated causal relationships. The overall causal retention heatmap of weight initialization method is given in Figure 4.6. From a performance perspective, among those experiments utilizing both the causal regularization and causal weight initialization techniques, the R-squared values peaked at 0.881. Furthermore, as depicted in Figure 4.5(d), models with causally initialized weights tend to converge faster than their unmodified counterparts. This trend is observed in both unmodified and causally regularized models. Additionally, the mean loss curve of causally regularized models lies above that of the causal weight initialized models, indicating that integrating causal weight initialization into the training process can slightly enhance the performance boundary. Thus, while our developed causal weight initialization method may not incorporate causal information into the model in the same manner as causal regularization, it establishes a substantial foundation for initiating causal embedded applications and significantly bolsters their efficiency and robustness.

Based on the insights garnered from our study, here are some recommendations:

(1) Combine causal regularization and weight initialization: These distinct techniques both contribute significant advantages. Causal regularization nudges the model to align with the causal prior, while causal weight initialization imparts an initial bias towards known causal relationships. The concurrent application of both techniques

can enhance the model's learning process, aiding in the capture and retention of causal information.

(2) Maintain balanced structural complexity: Avoiding structures that are either overly simplistic or excessively complex is key to robust performance. Striking the right balance ensures the model possesses the requisite complexity to capture causal relationships without succumbing to data overfitting.

(3) Consider activation function selection and adjust regularization strength: For causal priors that are approximately linear, the piecewise linear nature of the LeakyReLU or ReLU activation functions can better harmonize with the prior. It is also important to note that the choice of regularization strength (λ) can significantly impact the model's ability to retain causal information. This was evident in the case of LeakyReLU, where different regularization strengths led to varying levels of causal information retention.

These strategies provide a comprehensive guideline for explicitly incorporating experiment-extracted causal prior into neural networks.

4.4 Summary

This study delves into the innovative integration of prior knowledge from experiments and physics-based models with neural networks, focusing on metformin to showcase this methodology. Building upon this foundation, this chapter introduces a causal prior-embedded neural network framework that significantly enhances model interpretability,

demonstrating a methodical balance between utilizing extensive datasets and applying expert knowledge in environmental modelling and management. A causal and quantitative analysis of often-overlooked system parameters such as the Type-1 sorption fraction F along with first-order reaction rate coefficient α, and the scale of the transport system, was causally examined for the first time along with relevant confounders like particle density and saturation status. The effectiveness of the proposed methods has been thoroughly discussed and validated. The analysis of the experiment data, augmented data and the causal estimates overall showed that metformin's considerable long-range transport potential in porous media largely relies on its high relative velocity to water and extended half-life in groundwater. Such insight warrants a more comprehensive environmental assessment and increased public awareness about the risks of pharmaceuticals in the water cycle.

CHAPTER 5
A TRANSFER LEARNING APPROACH FOR MAPPING THE GLOBAL ENVIRONMENTAL RISK OF METFORMIN[4]

[4] This chapter is based on and expanded from the following paper:
Kang, Q., Yang, M., Song, X., Cao, Y., Liu, B., Ye, X., Wu, H., Chen, Y., Zhang, B. & Chen, B*. (2024). Mapping the Global Environmental Risk of Metformin: A Transfer Learning Approach. *Ready to submit.*
Roles: I conceived the study with the input from Chen, B, Yang, M. and Song, X. I designed the transfer learning-based framework and the EffluentNet structure. I developed the program scripts for processing, analyzing and visualizing the data along with Ye, X., and Wu, H. Visualization and interpretation of the results were further enhanced by the contributions from Yang, M., Cao, Y. and Song, X. The manuscript's refinement benefited from the comments and discussions from Chen, Y., Zhang, B., and Chen, B.

5.1 Introduction

Emerging pollutants refer to a group of pollutants that are not routinely monitored but have demonstrated potential environmental or ecological adverse effects (Ng et al., 2023; Y. Choi et al., 2021). Due to lower awareness, research on Contaminants of Emerging Concern (CECs) tends to commence more gradually compared to conventional pollutants that have garnered greater attention, despite the threats posed by many CECs are often comparable to those of more well-known pollutants(Wilkinson et al., 2022). Metformin is one such CEC. Since its introduction in the 1920s, the oral antihyperglycemic agent has been indispensable in managing type 2 diabetes mellitus. One in every two patients with type-2 diabetes is anticipated to be prescribed metformin (Drzewoski & Hanefeld, 2021; Ogurtsova et al., 2017). Beyond its primary role as an antidiabetic drug, metformin has been recognized for various additional effects, including potential life extension, anticancer properties, COVID-19 mitigation, and anti-mycobacterial benefits, in various studies (Böhme et al., 2020; Scheen, 2020; EL-Arabey & Abdalla, 2020; Romero et al., 2017; Dowling et al., 2012). Given these diverse reported benefits and the increasing global prevalence of type 2 diabetes, which is projected to reach approximately 580 million individuals worldwide by 2040, a continuous rise in the global consumption of metformin is anticipated (Ogurtsova et al., 2017). However, due to its limited metabolization in the human body, extensive usage, and inefficient removal by the secondary treatment technologies equipped in most contemporary wastewater treatment plants (WWTPs), metformin has become ubiquitously present in various aquatic compartments: based on Global Monitoring of Pharmaceuticals Project, the Joint Danube Survey 4, and multiple regional studies, metformin ranks among the most frequently detected PPCPs in surface

water, groundwater, and WWTP influents/effluents globally (Wilkinson et al., 2022; Zheng et al., 2023; Briones et al., 2018; Ambrosio-Albuquerque et al., 2021; R. Zhang et al., 2021; ICPDR, 2020; Tao et al., 2018; Oldenkamp et al., 2018; Yao et al., 2018; Briones et al., 2016; Trautwein et al., 2014). This widespread occurrence of metformin is further evidenced by its detection in potable water (Scheurer et al., 2012; R. Zhang et al., 2021). On the toxicity front, metformin has shown endocrine disruptive effects on aquatic life, including intersexuality, altered gene expression, and developmental changes, at concentrations as low as 40 μg/L(Elizalde-Velázquez & Gómez-Oliván, 2020; Niemuth & Klaper, 2018; MacLaren et al., 2018; Niemuth et al., 2015; Niemuth & Klaper, 2015). The toxicity of metformin's chlorination byproducts to human cells and living organisms also accentuates the environmental and health concerns associated with such widespread presence (R. Zhang et al., 2021; Y. He, Jin, et al., 2022). Consequently, understanding the global distribution of metformin in the environment is crucial for formulating effective regulations for metformin and other similar PPCPs, which are not only considered emerging pollutants but also biomarkers of human activity and epidemiological indicators (Wilkinson et al., 2022; Zheng et al., 2023; Shao et al., 2023; Lertxundi et al., 2023; Y. He, Zhang, et al., 2022).

In recent years, machine learning and deep learning methods have gained prominence in pollution risk analysis due to their ability to handle complex datasets effectively, deliver satisfactory performance, and remain cost-effective(Barzegar et al., 2018; Sajedi-Hosseini et al., 2018). These approaches have been applied in various global-scale assessments, contributing significantly to identifying pollution hotspots and facilitating more targeted policy development (Podgorski & Berg, 2020; Tang et al., 2021). However, when it comes

to metformin, a distinct challenge arises from its limited monitoring data. Unlike some other contaminants, routine metformin monitoring is not commonly included in environmental regulations in most places, resulting in insufficient data to map its global distribution comprehensively. Such a situation led to some efforts to estimate metformin occurrence using pharmaceutical consumption estimates, providing valuable insights, albeit constrained by data accessibility (Yang et al., 2022). This situation highlights the necessity for more specialized and innovative modelling approaches that can adeptly navigate the constraints of data scarcity, specifically in the context of metformin analysis (Ng et al., 2023; Wilkinson et al., 2022; Lertxundi et al., 2023; Bai et al., 2018). As a machine learning paradigm, transfer learning appears well-suited for addressing the challenges mentioned above. It involves applying knowledge from solving one problem to a different but related issue by leveraging pre-trained models based on extensive datasets(S. J. Pan & Yang, 2010). This is particularly beneficial when the available data for the new problem is sparse or needs to be more diverse. Its applications have spanned various studies, from evaluating groundwater quality to assessing air pollution levels, and demonstrated impressive results (W. Ma et al., 2022; Z. Chen et al., 2021; Hao et al., 2020; Fong et al., 2020). Thus, transfer learning was chosen as our primary methodology in this study. Additionally, it is crucial to develop a modelling strategy that effectively utilizes most of the relevant data from scattered sources for a comprehensive environmental analysis of metformin. Given the significant release of metformin into the environment through WWTPs, forecasting its concentrations in WWTPs worldwide is a pertinent and feasible approach supported by the availability of global WWTP datasets (Ehalt Macedo et al., 2021). By composing a dataset including metformin consumption, type-2 diabetes

prevalence, socioeconomic status, and the treatment capabilities of WWTPs, models that provide foundational insights for broader environmental impact assessments can be trained. This approach offers a detailed understanding of metformin's immediate occurrences around WWTPs and allows for an estimated global distribution, enhancing our comprehension of its broader ecological implications. This method could also prove invaluable for other PPCPs similar to metformin, particularly those strongly correlated to human activities.

This chapter proposes a strategy aimed at predicting metformin concentrations in WWTP influents/effluents and further estimating its environmental risks on a global scale, a task made challenging by the scarcity of comprehensive data. Our approach harnesses the potential of transfer learning to optimally utilize relevant, albeit distinct, background datasets, enabling effective learning from the limited data available on metformin's global occurrences (Cao et al., 2022; Z. Chen et al., 2021). A novel neural network architecture, EffluentNet, which appropriately considers endogenous causal relationships within the wastewater treatment process to enhance predictive performance, is introduced as a valuable addition to the toolkit for environmental analysis, particularly for tasks involving estimating pharmaceutical concentrations in wastewater. This study aims to estimate the global distribution of metformin, aimed at pinpointing areas with relatively higher potential risk. This effort aims to enable precise, targeted interventions and to enrich our understanding of the complex dynamics between type-2 diabetes prevalence, socioeconomic factors, wastewater treatment technologies, and metformin occurrences in global water. By achieving this, the study aspires to develop more effective and precise

management strategies for metformin and similar emerging pollutants, thereby contributing to better environmental stewardship and public health outcomes.

5.2 Methods

5.2.1 Modelling and Data Handling Strategies

Unlike ordinary pollutant with abundance reports in both urban and natural environments, the availability of real-world data concerning metformin concentrations in global WWTPs is limited and insufficient to build a robust estimator at a global scale. Thus, this chapter introduces a transfer learning strategy to mitigate such challenges. Transfer learning is valued for its capacity to utilize pre-existing datasets and models to address new, related challenges. And by its nature, it is particularly useful under a data-scarce context(Cao et al., 2022; Z. Chen et al., 2021). In this study, our transfer learning strategy comprises 1) a semi-synthetic dataset representing antidiabetic drug occurrences in Organisation for Economic Co-operation and Development (OECD) country WWTPs, 2) a real-world dataset derived from a comprehensive literature review on global metformin occurrences in WWTPs, 3) EffluentNet, a tailored neural network architecture designed to estimate contaminant occurrences in WWTP influent and effluent, and 4) a customized model fine-tuning approach.

To process the data for further analysis, a series of processes were conducted for various data sources used in this chapter. HydroWASTE is a global dataset detailing WWTP characteristics such as geographic coordinates, population served, and effluent flow rate (Ehalt Macedo et al., 2021). In this study, it is used as the major source of the WWTP information. For WWTPs that lack clear geographic coordinates, a thorough investigation

to ascertain their locations through public available information (e.g., the naming and coding) was undertaken. This sometimes includes reasonable inferences with information such as maps, theses, reports, and government/company websites. Once the geocoordinates of a WWTP were confirmed, the HydroWASTE database was used as a reference for any missing parameters, such as discharge flow rate, population equivalent, and highest treatment level. The corresponding HydroWASTE ID for a WWTP was included whenever possible. In the event of discrepancies between the data provided by authors/governments and HydroWASTE database, precedence was given to the former, considering their greater likelihood of being more reflective of actual conditions at the time being. To maintain the representativeness of our dataset, WWTP records that remain unidentified after the extensive research above were excluded. Similarly, WWTPs that cannot be confidently associated with a specific community of less than 5 million people were omitted despite known geolocations (i.e., when only country or metropolitan area data is available).

This study aimed to use the raw prevalence rate of diabetes among 20-year-old inhabitants in the WWTP region (including both Type-1 and Type-2, diagnosed, and undiagnosed cases) to represent at a specific time, similar to the comparable diabetes prevalence published by IDF. However, routine health surveys, including diabetes prevalence data, are not commonly available year-round in most of the countries globally. Thus, a series of guidelines were developed to ensure the available information is reasonably utilized in this study while not being over-adjusted. Some principles include:

(1) When studies report prevalences of diagnosed diabetes (for instance, those derived from insurance records or surveys inquiring about diabetes diagnoses), the figures were adjusted to include an estimate of the proportion of undiagnosed patients, thereby obtaining a comprehensive prevalence rate that encompasses both diagnosed and undiagnosed individuals. Suppose the study does not provide information on undiagnosed cases. In that case, this dataset is supplemented with data from relevant studies or with those acquired by applying the country-level ratio of undiagnosed diabetes patients as suggested by the International Diabetes Federation (IDF). On rare occasions, public announcements from health authorities may also be utilized if they are deemed as the most appropriate source of information. Additional adjustments are typically unnecessary for research that presents blood-sample test outcomes and includes a random selection of community members.

(2) In cases where only the prevalence of Type-2 diabetes is reported, this study typically does not adjust for the presence of Type-1 diabetes due to its relatively lower prevalence among adults globally compared to Type-2, unless otherwise mentioned.

(3) When prevalence data are only available for the entire population, this study adjusts these figures to reflect the prevalence among individuals over 20 years old, using age structure data from reliable sources. However, when only prevalence

figures starting from 15 years old are available, they are used as reported without modification.

(4) Given the choice between raw and age-standardized prevalence data within a region, this study opts for the raw figures whenever possible. This preference is because age-standardized rates are optional for machine learning models that operate on datasets with multiple attributes.

(5) If the preferred source for diabetes prevalence data differs by more than three years from the corresponding study on metformin occurrences, and there is credible information available to estimate diabetes prevalence trends within the relevant region, the prevalence will be recalibrated.

(6) Data specific to smaller regions, such as counties, provinces, or cities, that meet quality criteria are favoured over broader estimates like national or regional data. If no quality data for smaller units are available, country-level diabetes prevalence figures from the IDF will be used by default, as used in the semi-synthetic dataset, as a substitute.

5.2.2 Semi-synthetic Dataset: Antidiabetic Drug in OECD WWTPs

To construct a semi-synthetic dataset, this study integrated several data sources: 1) annual consumption data for antidiabetic drugs in OECD countries (OECD, 2021), 2) HydroWASTE as mentioned earlier, 3) country-level diabetes prevalence from the International Diabetes Federation (IDF; Magliano & Boyko, 2021), and 4) annual GDP per capita figures for OECD countries. The drug consumption data is in alignment with the

Anatomical Therapeutic Chemical (ATC) classification/Defined Daily Dose (DDD) system by the WHO, where 'A10' denotes the ATC code for antidiabetic drugs and DDD represents the standardized daily dosage (WHO, 2024).

The occurrence of antidiabetic drug in the influent and effluent of each OECD WWTP can be estimated through the following formula:

$$A10_{influent} = Coef_{excretion} \cdot \frac{A10_{country} \cdot Population_{wwtp}}{1000 \cdot Discharge_{wwtp}} \quad (5.1)$$

$$A10_{effluent} = (1 - Eff_{treatment}) \cdot A10_{influent} \quad (5.2)$$

Where $A10_{influent}$ and $A10_{effluent}$ represent the presences of antidiabetic drugs in unit WWTP influent and effluent flow ($DDD \cdot day/m^3$) respectively, $A10_{country}$ represents country-level consumption of antidiabetic drugs in DDD per thousand population, $Population_{wwtp}$ denotes the population served by the WWTP, and $Discharge_{wwtp}$ refers to the daily flow rate of the wastewater discharged by the WWTP. $Coef_{excretion}$, the estimated excretion coefficient, was set at 0.575, considering the market share and reported excretion rates of various diabetes drugs in OECD countries(Moura et al., 2021; Soppi et al., 2018). $Eff_{treatment}$, indicative of the treatment efficiency for antidiabetic drugs at each WWTP based on its highest treatment level, was assigned values of 0.1, 0.3, and 0.8 for primary, secondary, and advanced treatment levels, respectively (Balakrishnan et al., 2022; Briones et al., 2018; Scheurer et al., 2012).

IDF published prevalence in the population over 20 years old, inclusive of both diagnosed and undiagnosed cases of type 1 and type 2 diabetes, was utilized as the data source of diabetes prevalence. The time frame for the OECD drug consumption data, GDP per capita,

and IDF diabetes prevalence spans from 2010 to 2021/2022. A conditional logic was applied for aligning IDF diabetes prevalence such that if the year was less than 2016, the prevalence data from 2011 was used; otherwise, the 2021 data was applied.

The semi-synthetic dataset was constructed by merging the aforementioned data sources based on country and year, comprising approximately 430,000 records. The dataset is termed 'semi-synthetic' due to its composite nature, integrating assumptions about certain relationships, yet still largely based on published datasets and is considered to reflect realistic scenarios.

5.2.3 Dataset for Fine-tuning: Metformin in Global WWTPs

Alongside the synthetic data, a smaller dataset comprising reported metformin concentrations in various WWTPs was compiled. Based on 31 selected studies that report metformin occurrences in WWTP influents and effluents, this dataset provided a realistic and specific set of observations for model fine-tuning and validation (Yao et al., 2018; Shao et al., 2023; Zheng et al., 2023; Y. He, Zhang, et al., 2022; Scheurer et al., 2012; Asghar et al., 2018; S. Wang et al., 2022; Inarmal & Moodley, 2023; Ogunbanwo et al., 2022; S. Choi et al., 2022; Y. Choi et al., 2021; Cardini et al., 2021; Golovko et al., 2021; González-Gaya et al., 2021; Sadutto et al., 2021; Shao et al., 2021; X.-B. Song et al., 2020; Ju et al., 2019; Xiao et al., 2019; Yan et al., 2019; Alygizakis et al., 2019; Burns et al., 2018; Oertel et al., 2018; K. H. Nguyen, 2018; De Jesus Gaffney et al., 2017; Shraim et al., 2017; Archer et al., 2017; Carmona et al., 2017; Kot-Wasik et al., 2016; Estrada-Arriaga et al., 2016; van Nuijs et al., 2010). It encompasses key WWTP parameters, such as geographical coordinates, population served, and wastewater discharge rates. Most of those

parameters are from the corresponding studies, supplemented by publicly available information, and, when necessary, data from HydroWASTE database. Besides metformin concentrations in WWTPs, considerable effort was invested to enrich this dataset with accurate representations of the diabetes epidemiological and economic status proximate to the WWTPs. Specifically, this enhancement involved incorporating diabetes prevalence and GDP per capita data, gathered at the most granular level available, extending in some cases down to the district level (Al-Rubeaan et al., 2014; Boehme et al., 2015; Brunetti et al., 2022; Bruun-Rasmussen et al., 2020; Cerovečki & Švajda, 2021; De Mestral et al., 2020; Fang et al., 2022; Laranjo et al., 2016; Menéndez Torre et al., 2021; Motlhale & Ncayiyana, 2019; D. Nguyen et al., 2020; Sahadew et al., 2022; Tamayo et al., 2016; Topor-Madry et al., 2019; Uloko et al., 2018). Data sources for GDP per capita in the chapter is given in Appendix C.1. Diabetes prevalence figures were standardized as mentioned above to represent the prevalence among individuals over the age of 20 years.

5.2.4 EffluentNet: A Customized Neural Network for Estimating Contaminants in WWTP

In the quest to estimate contaminant levels in both influent and effluent of wastewater treatment facilities, traditional data-driven methods typically follow one of three approaches: 1) predicting influent concentrations and then applying a treatment coefficient reflective of the highest level of treatment at the facility to derive effluent concentrations; 2) creating separate models for influent and effluent predictions and linking them sequentially; or 3) utilizing standard algorithmic models and treating influent/effluent types as a simple binary variable. These conventional methods, while straightforward, tend to

oversimplify the treatment process or inadequately ignored the intricate, nonlinear dynamics between various predictive features, potentially compromising the accuracy and reliability of the outcomes.

To address these challenges, EffluentNet, a dual-pathway neural network architecture crafted for the estimation of contaminants in both influent and effluent in a single model process, was proposed in this chapter. EffluentNet manages dependencies within the dataset variables to provide more robust predictions. As illustrated in Figure 5.1, EffluentNet processes common variables, such as diabetes prevalence rates, GDP per capita, alongside WWTP operational parameters, such as treatment level variables, via parallel pathways. While treatment-specific features pass through a dedicated treatment embedding layer, other features are processed through shared layers. An effluent mask categorizes the data, directing influent records through an influent-dedicated output layer and concatenating influent embeddings with treatment embeddings for effluent data, subsequently passing through an effluent output layer. This dual-pathway approach acknowledges the logical connection between influent and effluent, captures complex dynamics between WWTP operational parameters, epidemic-economic variables, and influent/effluent concentration of the contaminant, and also mitigates error propagation, enhancing the reliability of the model for decision-making support.

5.2.5 Hyperparameter Optimization, Model Fine-tuning, and Uncertainty Handling

Modelling training in this study follows the following flow: a series of models with varying neural network architectures were developed using a semi-synthetic dataset, designated as the “base models.” These base models underwent further fine-tuning with real-world

metformin concentration data from global WWTPs while searching for their optimal hyperparameters through a series of hyperparameter optimization experiments.

The first task was to develop predictive models for antidiabetic drug occurrences in wastewater across OECD countries, with architectures amenable to subsequent fine-tuning. To identify suitable structures for the base models, a comprehensive series of machine learning experiments focusing on hyperparameter optimization was undertaken. Prior to experimentation, comprehensive preprocessing was applied to the datasets. Given the variation in feature magnitudes between the semi-synthetic and fine-tuning datasets, standardization was necessary. To mitigate the risk of data leakage, a StandardScaler was fitted using the training set of the synthetic dataset. This scaler was subsequently employed for standardizing the fine-tuning data, ensuring no data leakage and the consistency of the method. Furthermore, a log transformation (np.log1p) was employed to normalize right-skewed distributions, particularly notable in the target variable indicating antidiabetic drug levels in WWTPs. The semi-synthetic and fine-tuning datasets were partitioned using a consistent ratio: 60% of the data formed the training set, while the remaining 40% constituted the test set.

Given the dataset sizes—approximately 400,000 for the semi-synthetic data and 400 for the fine-tuning dataset—the chosen architecture needed to strike a balance between being overly complex and overly simplistic. Opting for a conservative approach, a classic wide-narrow configuration was selected. The structural guidelines established were as follows:

(1) The number of hidden layers was constrained to a minimum of two and a maximum of five. Consequently, including the input and output layers, the neural networks in this study comprised between four and seven total layers.

(2) The architecture adhered to a Hanoi-tower pattern, ensuring that no layer was wider than its predecessor. For example, if the second layer contained 16 nodes, the subsequent layer could not exceed this number.

(3) A cap was placed on the maximum number of nodes, set at 208 for this study, with the first layer required to have either 64 or 32 nodes.

These criteria narrowed the field to 83 structural candidates. The chosen architecture featured a sequence of layers activated by ReLU functions, trained using the Adam optimizer with mean squared error as the loss function. The R^2 score was the primary metric for evaluating performance. An early-stopping mechanism was implemented to prevent overfitting during fine-tuning, halting training when no further improvements were seen on the validation set. Under this setting, preliminary experiments were conducted with various batch sizes (32, 64, 256, 4096) and learning rates (0.0001, 0.0003, 0.001, 0.003) across a range of advanced GPUs (2080TI, 3080 MOBILE, RTX 4090, L40, H100) and based on the model performance in this experiment, a batch size of 32 and a learning rate of 0.001 were chosen for training base models.

For each of the 83 selected candidate structures, training was conducted across various hyperparameter combinations to comprehensively cover a broad spectrum of reasonable hyperparameters. The combinations include one or two treatment embedding layers with

3, 4, or 5 nodes; learning rates of 0.001 and 0.005; and the option of a dropout layer before the final output, with dropout rates of 0.5 or 0, and led to 1,992 base model runs.

For comparative analysis, a classic Multilayer Perceptron model was trained for each parameter combination without the treatment embedding. The inclusion of treatment embedding—a key feature of EffluentNet—was shown to enhance model performance, while accurately capturing the relationship between influent and effluent concentrations.

Fine-tuning in machine learning entails adapting a pre-trained model to a specific and often smaller dataset, leveraging its pre-existing information to enhance performance on a new task. This method can conserve computational resources and improve model robustness, particularly when the new task has limited data availability as the scenario of this chapter (Cao et al., 2022; Z. Chen et al., 2021).To help the fine-tuning process, a fine-tuning strategy, a variant of the previously reported G***radualUnfreeze,*** was developed**.** The idea is that during fine-tuning, the approach first freezes the initial layers of the network, training only the latter layers to adapt to the new data. The improvement is that instead of setting a fixed amount to unfreeze the next layer, it uses a more flexible mechanism to unfreeze the next layer when there is no performance increase for a certain number of epochs, similar to early stopping (Bengio, 2012). This strategy was chosen based on the hypothesis that the base layers captured generalizable features relevant across both datasets. In contrast, the latter layers needed to adapt to the specificities of the fine-tuning dataset.

A pseudo-code for the algorithm is as follows:

```
Class GradualUnfreezeCallback:
    Initialize with monitor metric and patience parameter.
    FreezeBeforeTraining(pl_module):
        # Initially, freeze all layers except the last fully connected layer.
        For each layer in pl_module except the last fully connected layer:
            Freeze the layer.
    FinetuneFunction(pl_module, current_epoch, optimizer):
        # Unfreeze layers based on the performance metric and patience parameter.
        If the monitored metric has not improved for 'patience' epochs:
            If no layers have been unfrozen yet:
                Unfreeze the last fully connected layer.
            Else:
                Unfreeze the next layer up in the network.
            Reset the wait counter.
    OnValidationEpochEnd(trainer, pl_module):
        # Evaluate the model's performance on the validation set.
        Get the current value of the monitored metric.
        If the metric has improved or is the first evaluation:
            Update the best score to the current metric value.
            Reset the wait counter.
        Else:
            Increment the wait counter.
```

After fine-tuning, stringent criteria were applied to select models for the ensemble, aiming to exclude those generating extreme or unrealistic predictions, thereby maintaining the ensemble's integrity for real-world applicability. The selection criteria included:

(1) Capping the maximum predicted influent concentration at 10,000 μg/l.

(2) Setting the minimum predicted influent concentration strictly above 0.

(3) Limiting the maximum predicted effluent concentration to 1,000 μg/l or less.

(4) Excluding models where the 75th percentile of effluent predictions equaled the maximum.

(5) Requiring the minimum predicted effluent concentration to be above 0.

These measures ensured the exclusion of models with extreme or non-physical predictions. Following this rigorous selection process, 20 models were chosen for each embedding depth (1, 2, 3), resulting in 60 candidates for the ensemble.

For the next step, confidence intervals for model predictions are estimated using a non-parametric bootstrapping technique. This involves repeatedly sampling with replacement from the set of forecasts, each contributing a mean value to a distribution of means. From this distribution, the confidence interval was calculated by identifying the bounds within which a specified percentage (usually 95%) of these means lie. This bootstrapping approach, robust and assumption-free, accounts for the inherent variability in the data and provides a reliable measure of the uncertainty associated with our model's predictions.

All dataset handling, modelling and analysis in this section were conducted using Python 3.9+. Our machine learning framework of choice was PyTorch Lightning version 2.1 paired with CUDA version 12.1. Model training, evaluation, and management of machine learning experiments were streamlined using MLflow version 2.8.

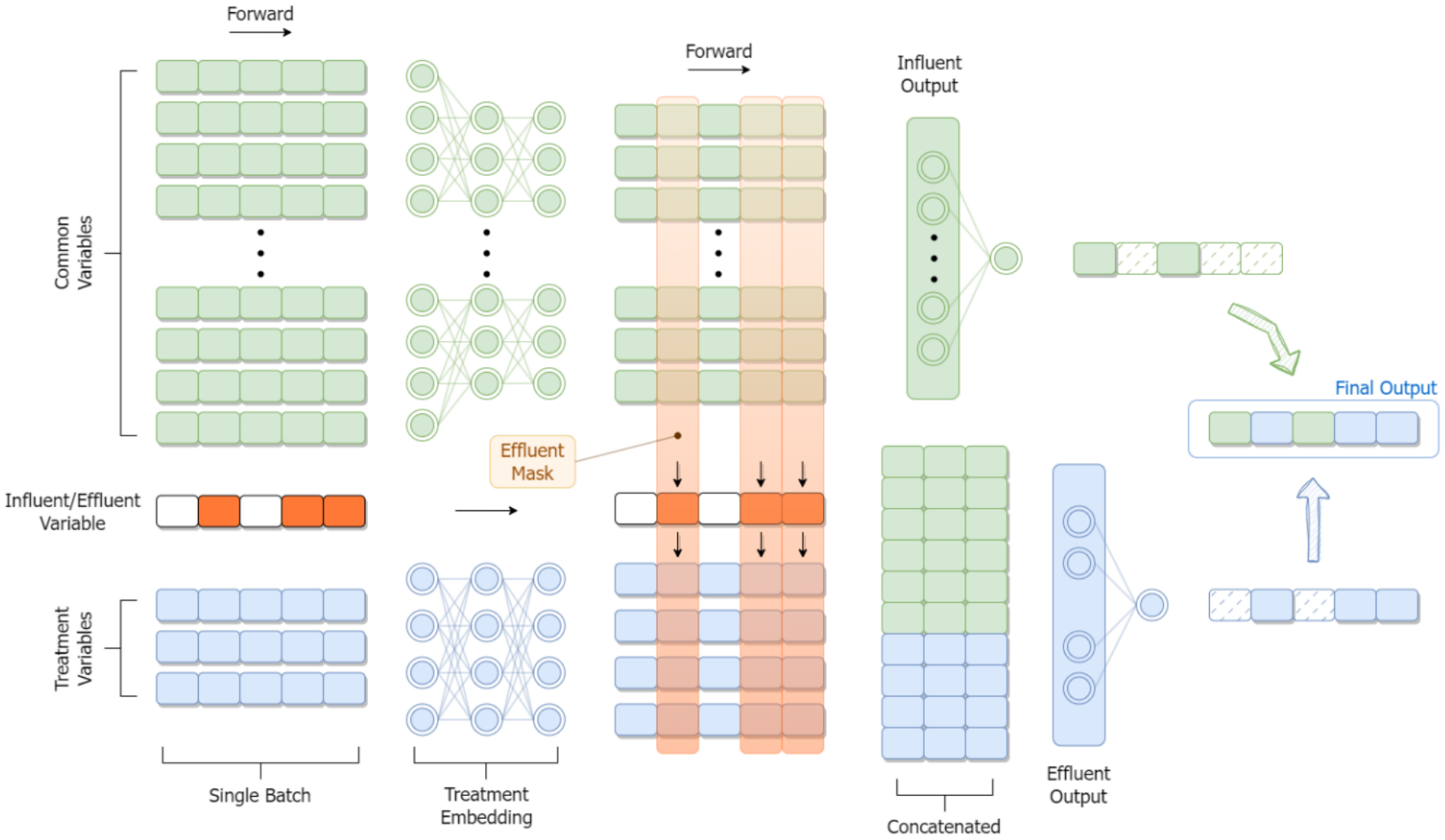


Figure 5.1 The neural network structure of EffluentNet.

5.3 Results and Discussion

5.3.1 Global Metformin Risk Quotients

To estimate the risk of metformin in a specific administrative unit with estimated metformin concentration in WWTPs, the Risk Quotient (RQ) for metformin at each WWTP was initially calculated:

$$RQ = \frac{C_{effluent}}{PNEC \times DF} \quad (5.1)$$

Where $C_{effluent}$ denotes the concentration of effluents from individual WWTPs, DF is the dilution factor for a WWTP, and $PNEC$ is the Predicted No-Effect Concentration (PNEC) for metformin, set at 160 µg/L based on the NORMAN database(Dulio et al., 2018). The definition of DF here is the same as in the HydroWASTE study and represents the ratio between the natural discharge of the receiving waterbody and the WWTP effluent discharge, which is given as:

$$DF = \frac{Q+W}{W} \quad (5.2)$$

Where Q is the receiving water's daily flowrate (m^3/day) and W is the flowrate of the discharge (m^3/day)(Ehalt Macedo et al., 2021).

Subsequently, the aggregate RQ in each region using the following formula:

$$RQ_{region} = \sum(RQ) \quad (5.3)$$

Here, RQ_{region} denotes the aggregate RQ in a specific administrative unit.

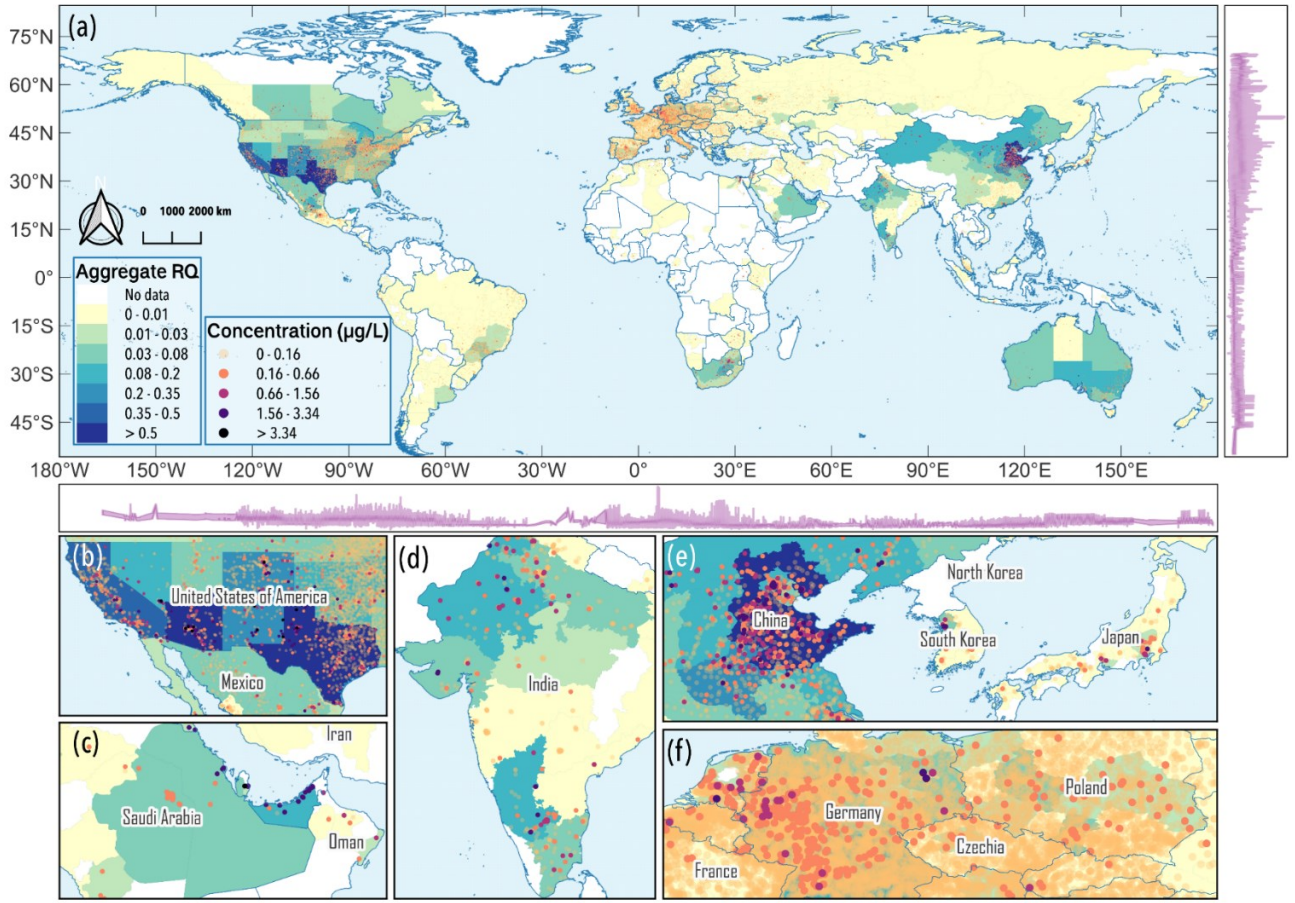


Figure 5.2 Global estimation of aggregate metformin Risk Quotients (RQ) and concentrations. Each dot signifies a WWTP discharge point, with the color reflecting the estimated metformin concentration in the vicinity of the discharge area. The RQ for each administrative unit is calculated as the sum of individual RQs within that unit. Subplots (b) to (f) detail specific regions: the Southwestern U.S., the Arabian Peninsula, India, East Asia, and Europe.

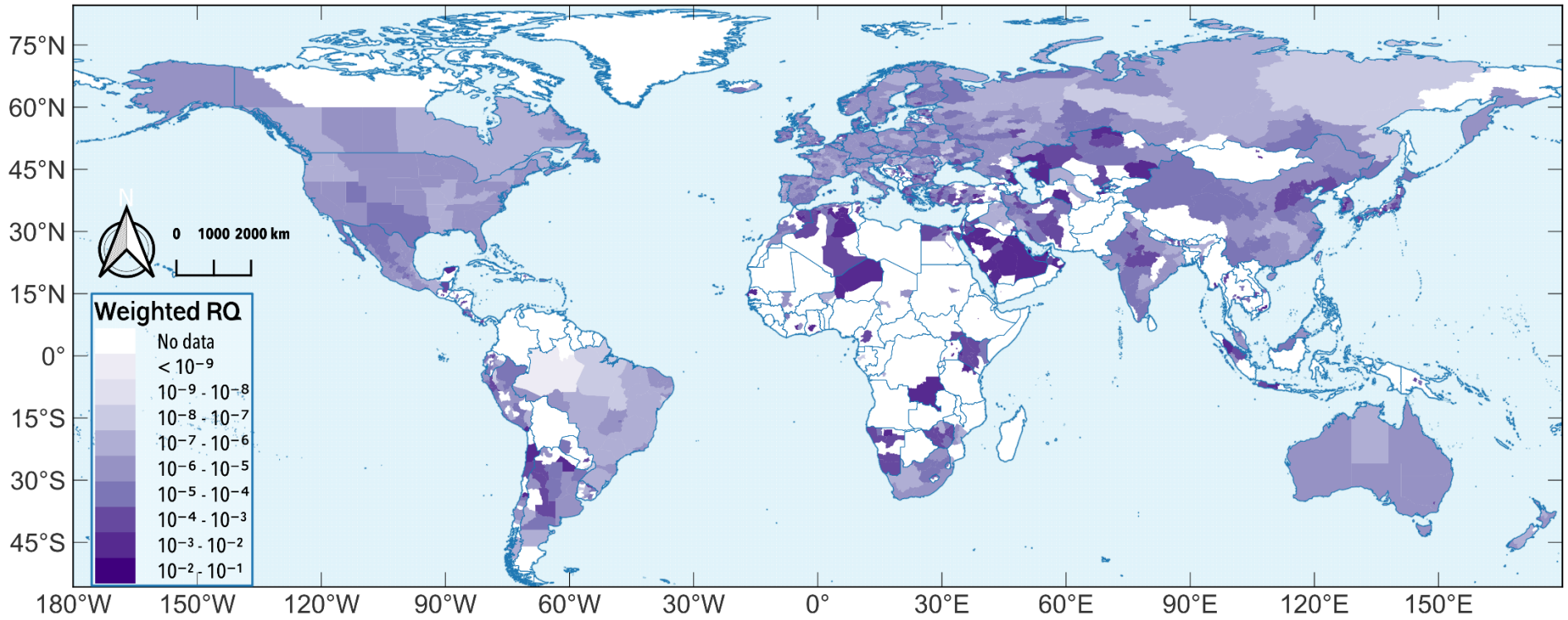


Figure 5.3 Global estimation of weighted average metformin Risk Quotients (RQ).

The spatial mapping of RQ values, derived from the estimated metformin concentrations, pinpointed regions potentially more vulnerable to metformin discharge into the environment while underscoring the heterogeneous nature of its risk distribution at a global scale, as depicted in Figure 5.2. The primary map, Figure 5.2(a), offers a global overview of metformin estimated risks. Confidence intervals, adjacent to Figure 5.2(a), visualized the uncertainty along both longitudinal and latitudinal axes, mirroring the variability inherent in our estimations. It can be seen that the north sphere shows apparent higher variability compared with the south sphere due to a higher data density field (Ehalt Macedo et al., 2021). The metformin RQ, while varying widely in magnitude, is predominantly low across most regions. However, specific regions exhibit heightened risk levels, portrayed with darker hues, signalling potential higher environmental stress, which can be attributed to metformin consumption and wastewater discharges. This pattern aligns with findings that suggest metformin's ubiquitous presence with over 50% of samples in numerous global and regional studies exceeding the Limits of Quantitation, while the reported concentrations spanning a wide range of magnitudes from low ng/L level to high μg/L level (Y. He, Zhang, et al., 2022; ICPDR, 2020; Ng et al., 2023; Shao et al., 2021; Wilkinson et al., 2022). Figure 5.2(b-f) demonstrated some regions with elevated metformin risks, particularly in the Southwestern United States, the Arabian Peninsula, Central Europe, the Indian subcontinent, and East Asia. The diverse socioeconomic and ecological characteristics of those regions underscore the intricate nature of metformin risks.

The region of the Southwestern United States, marked by its vast agricultural expanses and significant urban areas, is now depicted with an RQ indicative of potential concern as in Figure 5.2(b). Given the area's existing water stress, the heightened RQ values call for

improved wastewater management strategies to address the compounding effects of agricultural runoff, urban discharge, and limited water availability (Miller et al., 2021; Vörösmarty et al., 2000). Arabia Peninsula is another critical area with generally elevated RQ, as seen in Figure 5.2(c). In light of the notably rapid urbanization process and high prevalence of diabetes, compounded by the arid climate and acute water scarcity in the Arabian Peninsula, effluent management stands out as a paramount concern (Alotaibi et al., 2017; Ogurtsova et al., 2017). In Figure 5.2(d), the Indian subcontinent showcases a complex array of aggregate RQ levels for metformin, indicative of varying degrees of ecological risk across this densely populated and rapidly developing region. Similarly, the varied RQ profile in East Asia, as in Figure 5.2(e), also exhibits disparities akin to those observed in the Indian subcontinent, which presents a significant concern considering the region's densely populated urban centers and extensive agricultural lands (Y. Choi et al., 2021; Tanabe & Ramu, 2012; Yan et al., 2019). Notably, areas adjacent to the Yellow Sea—including China's Jing-Jin-Ji Metropolitan Region, Shandong and Liaoning provinces, and South Korea's Incheon—stand out as particularly impacted zones and show clear signs of considerable metformin stress. The proximity of these high-risk zones in Asia to significant river systems, such as the Ganges, Yellow River and Han River, might also contribute to the dissemination of metformin in the area and require further investigation. In Europe, Germany and its neighbouring countries present an elevated risk profile, with certain zones showing increased aggregate RQ, as depicted in Figure 5.2(f), which suggests a potential for metformin to impact not only freshwater systems but also to carry over into the North Sea and Baltic Sea. Meanwhile, some other European regions like the Iberian Peninsula and areas surrounding the Black Sea also exhibit slightly higher RQs. These

observations underscore the need for more collaborations to tackle the emerging pollutants span beyond national borders (Dulio et al., 2018).

In general, the variance in RQs across the globe mirrors the differing degrees of industrialization, the efficacy of wastewater treatment processes, and the prevalence of diabetes (Ehalt Macedo et al., 2021; Ogurtsova et al., 2017). The heightened RQ levels in the key areas reflect the substantial use of metformin, stemming from a considerable prevalence of diabetes, compounded by the challenges of managing the effluent outputs from urban centers that often lack advanced wastewater treatment. A complementary global metformin risk map that utilizes weighted average RQ for each region is provided for comparison in Figure 5.3, offering parallel insights into the spatial distribution of metformin risk profile.

5.3.2 Metformin in Canadian Ecozones

Canada's diverse landscapes, ecological environments, coupled with a spectrum of human interactions ranging from the indigenous communities utilizing traditional food sources to the dynamic urban lifestyles in densely populated areas, presents an ideal background for a comprehensive, zoomed-in risk assessment of metformin and similar PPCPs. Its sensitivities among subarctic and coastal regions also warrant close attention (Drever et al., 2021; Sanborn et al., 2011). The sparse data available in Canada presents a unique opportunity to test and validate our transfer learning-based approach (Littlejohn et al., 2023; Schwartz et al., 2021; Ghoshdastidar et al., 2015). Given the reasons outlined above, a detailed exploration of metformin's risk within the Canadian context was undertaken.

Our estimated figures for average metformin concentrations in influents and effluents across 1,710 WWTPs in Canada closely align with the metformin concentrations reported in various studies. These existing reports indicate metformin concentration levels in Canadian effluents varying at low μg/L levels, specifically 0.472 to 10.6 μg/L in Nova Scotia, 70 μg/L in Hamilton, Ontario, and 3.6 ± 3 μg/L in North Bay, Ontario(Ghoshdastidar et al., 2015; Littlejohn et al., 2023; Parrott et al., 2021). It is particularly noteworthy that our methodology deliberately excluded data from Canada in both the background dataset for base model training and the real-world dataset for fine-tuning, specifically to ensure the integrity of our validation process free from data leakage in any form. This strategic approach underscores the robustness and practicality of our transfer learning technique, as our results falling within the reported range of Canadian metformin effluent concentration, despite its exclusion from our initial datasets. Moreover, the training metrics suggested that our model ensemble explained over 60% of the variance in the global metformin WWTP occurrences dataset (average R^2:0.63), indicates that our season-neutral estimates can offer valuable insights and a reliable basis for broader global assessments and affirms the potential utility for more expansive risk assessment.

The Canadian Ecozone system offers a classification that reflects natural ecological divisions, essential for a precise assessment of metformin's environmental impact, and it intrinsically aligns with the First Nation communities' traditional land use and dietary practices(Marshall et al., 1996). Thus, the RQ in each Canadian ecozone was calculated using equation (5.3) with the results presented in Figure 5.4(a). Overall, Canada's risk from metformin is found to be low. This finding is consistent with the reported 21% detection rate of metformin in surface water samples across 11 ecozones, from First Nations Food,

Nutrition & Environment Study (FNFNES), a decade-long survey aims to promote healthy environments and healthy food for healthy First Nations(H. M. Chan et al., 2021; L. Chan et al., 2019). Nonetheless, there is a noticeable variation in risk across different ecozones. Regions such as the Prairies, Mixedwood Plains, and Boreal Plains are identified as higher risk areas, correlating with regions of heightened human activity. In contrast, metformin risk in ecozones such as Boreal Cordillera, Taiga Plains, and Taiga Shield are notably lower, primarily due to minimal recorded point source discharge. These results also correlate with the varied detection rates in these ecozones: during the survey, Boreal Cordillera and Taiga Plains reported no metformin detection while samples from Taiga Shield got a low metformin detect rate at 1/15, respectively, whereas Mixedwood Plains and Boreal Plains exhibited detection rates of 24/24 and 6/54, with the highest concentrations in the samples being 2020 ng/L and 93 ng/L. The alignment of our findings with reported detection rates is consistent across most ecozones except for Prairies, where our estimation indicates a higher risk, contrasting with a lower detection rate of 1/18 from the survey. This discrepancy could be attributed to limited sample locations in the survey, despite very high human activity in the region. Among all the ecozones, the Mixedwood Plains, as Canada's most densely populated and industrially active ecozone, faces substantial PPCP contamination risks. The close interactions in between urban centers and prime agricultural lands, coupled with the ecozone's rich waterways, create a complex challenge in managing PPCP pollution in the Mixedwood Plains, which necessitate advanced wastewater treatment solutions to protect the water quality and public health (Chambers et al., 2012). The Prairies, characterized by extensive agricultural activities, also represent a significant concern for metformin dispersion due to the vast farmlands that

cover more than 90% of the land base. The extensive use of water for irrigation and the resultant runoff make this ecozone particularly vulnerable to PPCP contamination, potentially impacting both the local biodiversity and human health through the consumption of contaminated water and food sources (Marshall et al., 1996; Bartzen et al., 2010). While less dominated by agriculture, with only about 20% of the land devoted to farming, the Boreal Plains are not immune to risks. The ecozone's significant forestry industry and emerging oil and gas development introduce various pathways for PPCP infiltration into aquatic systems. The extensive network of rivers and lakes in this area could also facilitate the spread of contaminants, affecting both aquatic life and the communities reliant on these water sources (Ireson et al., 2015).

Fishing is recognized as the predominant food harvesting activity at the household level among First Nations communities across all ecozones, as reported by FNFNES (H. M. Chan et al., 2021; L. Chan et al., 2019). In light of this, this study seeks to identify ecozones where the consumption of natural products intersects with a non-negligible risk of metformin exposure, which aims to inform future environmental monitoring, public health policies, and community awareness programs, ensuring a culturally sensitive approach to environmental management (Schwartz et al., 2021). Applying this criterion, it is noticeable that the Boreal Shield, a region renowned for its iconic Canadian wildlife, spans over 1.8 million square kilometers and boasts substantial freshwater resources, presents a moderate risk profile for metformin while ranked sixth in average daily consumption of traditional food in the First Nation communities. The Montane Cordillera and Pacific Maritime ecozones, known for their ecological diversity and significant agricultural and forestry industries, ranking second and third respectively for average daily traditional food

consumption, exhibit low metformin risk profiles. In these ecozones, fish species such as trout (Oncorhynchus mykiss), walleye (Sander vitreus), and eulachon (Thaleichthys pacificus) are commonly consumed as traditional food. Considering the progressingly frequent reported ecotoxicity of metformin to aquatic lives (MacLaren et al., 2018; Niemuth et al., 2015; Y. He, Zhang, et al., 2022; Lin et al., 2021; Ussery et al., 2019; Jacob et al., 2019; Caldwell et al., 2019; Godoy et al., 2018; Markiewicz et al., 2017; Melvin et al., 2017), as well as the fact that metformin may serve as an indicator for the presence of various other PPCPs or trace contaminants (Y. He, Zhang, et al., 2022; Wilkinson et al., 2022), the ecozones mentioned above may require further investigation and research efforts to estimate the presence and bioaccumulation potential of metformin and other emerging pollutants, not only for safeguarding the health and well-being of their communities amidst ongoing environmental challenges but also for preserving the ecological integrity and cultural heritage of these regions.

5.3.3 Metformin in Arctic and sub-Arctic Regions

The Arctic and sub-Arctic regions, characterized by their pristine environments, unique biodiversity, and the integral role of indigenous communities, face significant challenges in PPCP pollution. The distinct climatic conditions and the relative isolation of these areas pose unique challenges for wastewater management, potentially exacerbating the risks associated with pharmaceutical contaminants. Thus, Figure 5.4(b) demonstrates the known locations of 85 WWTPs situated within or near the Arctic Circle (66° 34' N) where estimated concentrations within the effluent vicinity exceed the Limit of Quantification, set at 10 ng/L. WWTPs within 100 kilometers of each other are clustered to illustrate the potential zones of influence. The map illustrates that Sweden, Finland, and Russia have

WWTPs located directly within the Arctic Circle with estimate metformin discharge at concentrations above the detection limit. The United States, Canada, and Norway also operate several WWTPs with higher metformin effluent concentrations near the Arctic Circle above the 60th parallel north. Those facilities collectively discharge approximately 553,000 m^3 of wastewater per day into natural water bodies, with an estimated average metformin effluent concentration of 2.76 μg/L. Such a concentration is non-negligible for the Arctic regions. Taking the likelihood of metformin and other compounds associated with human activity affecting local food webs and indigenous communities (Chaves-Barquero et al., 2016), and the fact that effective wastewater treatment systems are rarely established in communities in the Arctic region (Gunnarsdóttir et al., 2013), it is reasonable to assume that the discharge of emerging pollutants, especially PPCPs like metformin, could have far-reaching effects beyond the immediate vicinity of the WWTPs such as potential long-range transport in natural water, as well as bioaccumulation and biomagnification in Arctic food webs. Thus, more advanced and nuanced management programs are needed to protect the sensitivity of Arctic ecosystems.

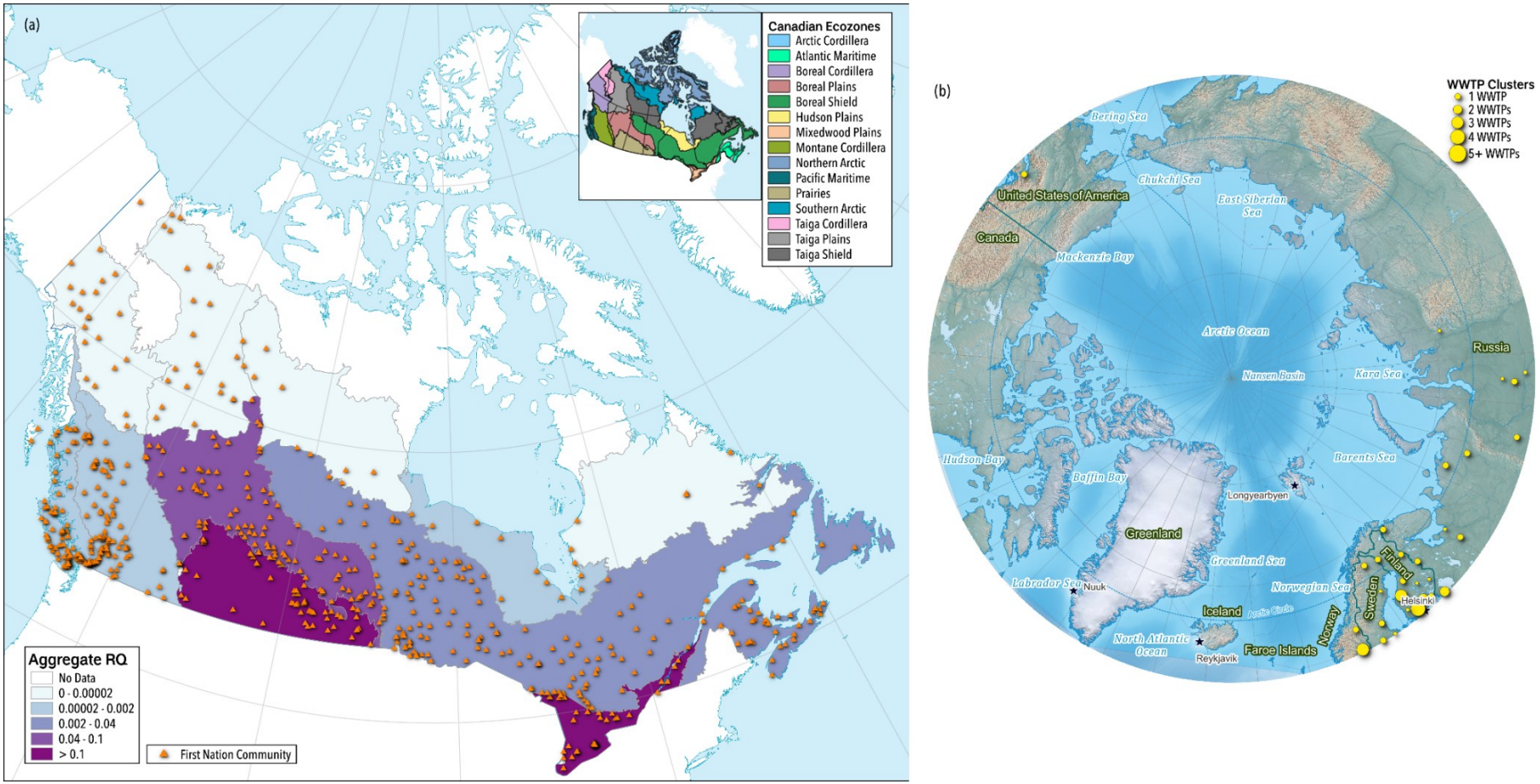


Figure 5.4 Metformin Risk in Canadian Ecozones and sub-Arctic/Arctic Regions. (a) Aggregate Metformin Risk Quotient (RQ) Estimations in Canadian Ecozones and Locations of First Nation Communities; (b) Estimated Metformin Occurrences Exceeding the Limit of Quantification (LoQ) of 10 ng/L in the Vicinity of WWTPs in Arctic and Subarctic Regions.

5.4 Summary

This chapter introduces a scalable and efficient model for environmental risk assessment that merges environmental engineering with cutting-edge data science techniques, complemented by a series of innovative techniques to aid the method including a new neural network architecture EffluentNet, and a data fine-tuning strategy. The ecological risk of metformin is estimated with the framework, as a case study. By employing transfer learning, augmented with a limited set of domain-specific data, the extensive potential of this approach in applied environmental research and its utility in enhancing the robustness and relevance of our risk assessments have been demonstrated, particularly in data-sparse regions. This is further validated by our model's alignment with observed metformin levels in Canada. This study offers a comprehensive evaluation of the ecological risk posed by metformin across diverse global regions, employing spatial mapping to visualize metformin discharges based on existing data. The investigation into the widespread distribution of metformin across aquatic systems worldwide underscores the urgent requirement for strategic policy responses, further highlighted by the significant spatial variability in risk levels, intricately tied to a combination of socioeconomic, industrial, and ecological factors. Our case studies for the Canadian ecozones, the Arctic and sub-Arctic regions, draw attention to the concern of pharmaceutical pollutants encroaching upon these pristine environments.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Conclusions

The overarching goal of this dissertation research is to enhance environmental data analysis by introducing innovative data analysis frameworks and methodologies, which aims to deepen our comprehension of environmental dynamics, addressing both the complexity of environmental systems and the intricacy of environmental data. The dissertation presents the developed methodologies and frameworks that encapsulate the integration of causal inference, physics-informed neural networks, and transfer learning techniques to tackle the multifaceted challenges in environmental data analysis. This goal serves as a foundation for the case studies detailed herein, each targeting specific aspects of environmental dynamics and offering unique insights into pressing environmental concerns. The key findings and contributions of the dissertation are as follows.

Chapter 3 demonstrates a new causal reasoning method based on observational data with the aid of causal inference models and machine learning techniques. To investigate a causal problem with observational data, prior knowledge as an indispensable part of the system was also considered. In the case study on the interrelation of COVID-19 and air pollution, the socio-economic and temporal factors information was brought into the equation by explicitly identifying interrelations between variables in the directed acyclic graph and slicing the data through multiple data processing techniques such as city clustering and phase-wise analysis. Through the observational data from 166 Chinese cities, most of the reported potential causal relationships between environmental factors and COVID-19 severity from a short-term perspective with the proposed causal inference framework were examined. Based on the results, most of the estimations of the links (89 out of 90) under nine different cluster-phase settings were refuted. The results showed that the impact caused by environmental factors on the severity of COVID-19 was limited across

all three clusters. Commonly discussed factors such as rational policymaking, sufficient public awareness, and effective isolation strategy are still crucial for containing the ongoing COVID-19 pandemic.

Chapter 4 explores the potential of integrating prior knowledge extracted from experiments and physics-based models into neural networks using metformin as a case study for demonstration. Underexplored system parameters such as the type-1 sorption fraction F, first-order reaction rate coefficient α, and transport system scale have been causally and quantitatively evaluated with adequate confounders considered. The analysis of the experiment data, augmented data and the causal estimates overall showed that metformin's considerable long-range transport potential in porous media largely relies on its high relative velocity to water and extended half-life in groundwater. Such insight warrants a more comprehensive environmental assessment and increased public awareness about the risks of pharmaceuticals in the water cycle.

Chapter 5 developed a novel modelling approach for assessing the risks posed by emerging pollutants with limited data availability, including a neural network architecture EffluentNet for estimating the occurrences of the water within influent and effluent or similar distributions, and a data fine-tuning strategy to maximally utilize available data. The ecological risk of metformin is estimated with the framework, as a case study. Metformin's global risk is estimated for the first time, providing important value in environmental policy making. The investigation into the widespread distribution of metformin across aquatic systems worldwide underscores the urgent requirement for strategic policy responses, further highlighted by the significant spatial variability in risk levels intricately tied to a combination of socioeconomic, industrial, and ecological factors. The case studies for the Canadian ecozones, the Arctic and sub-Arctic regions, draw attention to the concern of pharmaceutical pollutants encroaching upon these pristine environments. This

chapter also showcased a scalable and cost-effective modelling approach for environmental risk assessment and emerging pollutant management by merging environmental engineering and science with advanced data science techniques.

6.2 Research Contributions

This dissertation research contributes significantly to environmental engineering and science. The key research contributions are as follows:

Development of a Machine Learning-Aided Causal Inference Framework: This dissertation research successfully established an advanced framework incorporating interpretable machine learning techniques and causal inference methodologies. By recognizing the critical role of prior knowledge in unraveling causal problems, the study intricately weaved socio-economic and temporal considerations into the analysis. This was achieved by identifying variable interrelations within DAG and employing sophisticated data processing methods, including city clustering and phase-wise analysis. This approach could significantly enhance model interpretability and ensures causal insights can be extracted from observational datasets across environmental engineering and science studies, setting a new standard for clarity and accessibility in environmental models. The dissertation research also provided an in-depth causal analysis of various environmental factors, including air pollution and meteorology, on the severity of COVID-19. Utilizing a novel data analysis framework, this work illuminated the minimal causal impact these factors have on disease severity, thus redirecting focus towards more influential containment strategies. This aspect of the research distinguishes between spurious associations and genuine causal relationships, offering valuable insights for pandemic policy and response.

Development of a Casual Prior-Embedded Physics-Informed Neural Network Framework:

This dissertation research showcased the innovative integration of prior knowledge, derived from experimental and physics-based models with causal inference analysis, into neural networks. For the first time, this approach enabled the causal and quantitative evaluation of previously underexplored porous media transport system parameters taking into account critical confounders like particle density and saturation status. Metformin's environmental behaviour was investigated as a case study. This methodology provides a balanced approach to environmental modelling and management, navigating between the complexities of extensive datasets and the invaluable application of expert knowledge, ensuring the preservation of most physics-causal connections, merging data-driven insights with fundamental physical and causal principles without sacrificing analytical performance. This novel paradigm promotes a substantial improvement in both the interpretability and efficacy of AI and ML applications within the realms of environmental science and engineering. The dissertation research also delved into the under-researched area of the environmental fate and transport of metformin as a pharmaceutical pollutant. Through a combination of experimental data integration, physics-based modelling, causal inference, and neural networks, the study revealed critical insights into metformin's long-range transport potential. This highlights the necessity for a broader environmental risk assessment concerning pharmaceuticals in the water cycle.

Development of a Transfer Learning-based Environmental Risk Estimation Framework:

This dissertation research signified a methodological innovation by successfully integrating transfer learning techniques with environmental data analysis, specifically tailored to address the challenges of estimating the risks of emerging pollutants in diverse global regions. Additionally, the creation of EffluentNet and a novel fine-tuning method mechanism represent notable

exploration and advancements in adapting data science techniques. Those innovations ensure that the predictive models are not only grounded in empirical data but also reflect the underlying physical processes and causal relationships, making the outcomes more meaningful and actionable for environmental risk management. This methodological innovation seamlessly merges principles of environmental engineering and science with cutting-edge data science techniques, offering a robust and scalable solution for tackling data scarcity when investigating emerging problems. This dissertation research also offers a comprehensive global assessment of the environmental risk of metformin. Employing spatial mapping and transfer learning techniques, the dissertation identified significant variability in risk levels across different regions, focusing on Canadian ecozones and the Arctic and sub-Arctic regions.

6.3 Recommendations for Future Work

The following recommendations for future work are outlined to extend the boundaries of current knowledge in environmental data analysis.

1) **Enhancing Data Accessibility and Modelling Interpretability:** Future data-driven studies in the field should promote data accessibility and focus on model interpretability to increase the data for emerging environmental issues. These are always sufficient to cultivate high-quality research output and accelerate understanding of the problem.

2) **Embedding Complex Causal Priors and Novel Machine Learning Techniques:** Future work can explore embedding more complex causal priors into state-of-the-art machine learning algorithms, such as Graphical Neural Networks (GNNs). This avenue holds the potential to significantly enhance model interpretability and robustness, facilitating the application of AI in scientific inquiries with a causal perspective.

3) **Expanding Methodological Applications and Insights in Environmental Research** The methodologies developed and tested in this dissertation research demonstrated versatility and applicability across a broad spectrum of environmental contexts and challenges. Future research can further advance and extend these methods to address a wider range of environmental issues, such as non-point source pollution and seasonal variability. This would entail expanding the scope of analysis beyond WWTP effluent as showcased in the dissertation to include other environmental compartments, such as groundwater systems, and incorporating factors like runoff, non-point sources, and seasonal changes to achieve a more comprehensive and accurate risk assessment of pharmaceuticals like metformin. Additionally, the dissertation underscores the importance of advancements in wastewater treatment technologies. By developing and disseminating water treatment technologies that are both effective and economically viable, it is possible to manage the presence of emerging pollutants more effectively. Customizing these technologies to meet the specific needs and economic conditions of diverse communities worldwide will be crucial in ensuring universal access to clean water. This approach not only tackles the direct challenges posed by pharmaceuticals and other pollutants but also contributes to a broader strategy for sustainable water management and environmental protection.

4) **Robust Regulatory Frameworks and Community Engagement Incorporating Indigenous Knowledge:** Establishing scientific effluent standards and medication return programs, alongside educating communities about responsible pharmaceutical disposal, can significantly contribute to reducing emerging environmental contamination. Emphasizing diabetes prevention and management can also mitigate the prevalence of metformin in water systems. Incorporating the ecological knowledge and practices of indigenous communities and fostering international cooperation for addressing pharmaceutical pollution is imperative. efforts are crucial

for achieving sustainable environmental stewardship and ensuring the health and well-being of current and future generations.

5) **Cultivate Optimal Data Curation Practices:** The application of data-driven methodologies such as machine learning within the realm of environmental engineering reveals structural challenges that impede advancing research in this critical field. Firstly, the practice of employing machine learning techniques remains largely diverse, no cohesive set of protocols or best practices exists, making it challenging for researchers to apply and compare methodologies across different studies consistently. Additionally, the practice of data and model sharing in environmental engineering and science is far from ideal. Sharing the original data is still not considered a common practice in the field despite being encouraged by many peer-reviewed journals. To overcome these obstacles and foster a culture of innovation and collaboration in environmental engineering and science, it is imperative to cultivate optimal data curation practices and promote the sharing of data and models within the community whenever possible.

6.4 Selected Publications

During the Ph.D. program, I published 15 peer-reviewed journal articles and one conference proceeding paper. I was the first author on four of these papers and served as an equal contributor on two. Among my two unpublished works, one is currently under revision, and the other is prepared and ready for submission. In the remaining publications where I am listed as an author, I made substantial contributions to data analysis, encompassing method design, coding, visualization, and results interpretation.

Peer-reviewed Key Papers[5]

Kang, Q., Song, X., Xin, X., Chen, B.*, Chen, Y., Ye, X., & Zhang, B. (2021). Machine Learning-Aided Causal Inference Framework for Environmental Data Analysis: A COVID-19 Case Study. Environmental Science & Technology, 55(19), 13400-13410.

Kang, Q., Zhang, B., Cao, Y., Song, X., Ye, X., Li, X., Wu, H., Chen, Y. & Chen, B.* (2024). Causal Prior-Embedded Physics-Informed Neural Networks and a Case Study on Metformin Transport in Porous Media, *Water Research,* Accepted.

Kang, Q., Yang, M., Song, X., Cao, Y., Liu, B., Ye, X., Wu, H., Zhang, B. & Chen, B.*(2024). Mapping the Global Environmental Risk of Metformin: A Transfer Learning Approach. *Ready to Submit*.

Datta, A. R., **Kang, Q.,** Chen, B.*, & Ye, X. (2018). Fate and transport modelling of emerging pollutants from watersheds to oceans: a review. Advances in Marine Biology, 81, 97-128.

Peer-reviewed Collaborative Publications

Cao, Y.#., **Kang, Q.** #**,** Zhang, B., Zhu, Z., Dong, G., Cai, Q., Lee, K. & Chen, B.* (2022). Machine learning-aided causal inference for unravelling chemical dispersant and salinity effects on crude oil biodegradation. Bioresource Technology, 345, 126468. https://doi.org/10.1016/j.biortech.2021.126468

Fu, H., **Kang, Q.,** Sun, X., Liu, W., Li, Y., Chen, B., Zhang, B., & Bao, M.* (2023). Mechanism of Nearshore Sediment-Facilitated Oil Transport: New Insights from Causal Inference Analysis. Journal of Hazardous Materials, 133187. https://doi.org/10.1016/j.jhazmat.2023.133187

[5] # indicates equal contributors.

Yang, M., Zhang, B., Chen, X., **Kang, Q.,** Gao, B., Lee, K., & Chen, B.* (2023). Transport of Microplastic and Dispersed Oil Co-contaminants in the Marine Environment. Environmental Science & Technology, 57(14), 5633-5645. https://pubs.acs.org/doi/abs/10.1021/acs.est.2c08716

Chen, Y., Zhang, B., Yang, M., Xin, X., Kang, Q., Ye, X., & Chen, B.* (2022). An integrated framework of optimized learning networks for classifying oil-mixed microplastics. Journal of Cleaner Production, 379, 134698.

Chen, Y., Chen, B.*, Song, X., **Kang, Q.,** Ye, X., & Zhang, B. (2021). A data-driven binary-classification framework for oil fingerprinting analysis. Environmental Research, 111454.

Ye, X., Chen, B.*, Lee, K., Storesund, R., Li, P., **Kang, Q.,** & Zhang, B. (2021). An emergency response system by dynamic simulation and enhanced particle swarm optimization and application for a marine oil spill accident. Journal of Cleaner Production, 297, 126591.

Conference Proceedings

Kang, Q., Datta, A., & Chen, B.* (2021, May). Parameter Analysis in Simulating Transport of Metformin in a Sandy Medium. In *Canadian Society of Civil Engineering Annual Conference* (pp. 419-423). Singapore: Springer Nature Singapore.

BIBLIOGRAPHY

A, R. J. van der, Eskes, H. J., Boersma, K. F., Noije, T. P. C. van, Roozendael, M. V., Smedt, I. D., Peters, D. H. M. U., & Meijer, E. W. (2008). Trends, seasonal variability and dominant NOx source derived from a ten year record of NO2 measured from space. Journal of Geophysical Research: Atmospheres, 113(D4). https://doi.org/10.1029/2007JD009021

Accarino, G., Lorenzetti, S., & Aloisio, G. (2021). Assessing correlations between short-term exposure to atmospheric pollutants and COVID-19 spread in all Italian territorial areas. Environmental Pollution, 268, 115714. https://doi.org/10.1016/j.envpol.2020.115714

Adams, M. D. (2020). Air pollution in Ontario, Canada during the COVID-19 State of Emergency. Science of The Total Environment, 742, 140516. https://doi.org/10.1016/j.scitotenv.2020.140516

Ahmed, M. B., Zhou, J. L., Ngo, H. H., Guo, W., Thomaidis, N. S., & Xu, J. (2017). Progress in the biological and chemical treatment technologies for emerging contaminant removal from wastewater: A critical review. Journal of Hazardous Materials, 323, 274–298. https://doi.org/10.1016/j.jhazmat.2016.04.045

Alotaibi, A., Perry, L., Gholizadeh, L., & Al-Ganmi, A. (2017). Incidence and prevalence rates of diabetes mellitus in Saudi Arabia: An overview. Journal of Epidemiology and Global Health, 7(4), 211. https://doi.org/10.1016/j.jegh.2017.10.001

Al-Rubeaan, K., Al-Manaa, H., Khoja, T., Ahmad, N., Al-Sharqawi, A., Siddiqui, K., AlNaqeb, D., Aburisheh, K., Youssef, A., Al-Batil, A., Al-Otaibi, M., & Al Ghamdi, A. (2014). The Saudi Abnormal Glucose Metabolism and Diabetes Impact Study (SAUDI-DM). Annals of Saudi Medicine, 34(6), 465–475. https://doi.org/10.5144/0256-4947.2014.465

Alygizakis, N. A., Besselink, H., Paulus, G. K., Oswald, P., Hornstra, L. M., Oswaldova, M., Medema, G., Thomaidis, N. S., Behnisch, P. A., & Slobodnik, J. (2019). Characterization of wastewater effluents in the Danube River Basin with chemical screening, in vitro bioassays and antibiotic resistant genes analysis. Environment International, 127, 420–429. https://doi.org/10.1016/j.envint.2019.03.060

Ambrosio-Albuquerque, E. P., Cusioli, L. F., Bergamasco, R., Sinópolis Gigliolli, A. A., Lupepsa, L., Paupitz, B. R., Barbieri, P. A., Borin-Carvalho, L. A., & de

Brito Portela-Castro, A. L. (2021). Metformin environmental exposure: A systematic review. Environmental Toxicology and Pharmacology, 83, 103588. https://doi.org/10.1016/j.etap.2021.103588

Andree, B. P. J. (2020). Incidence of COVID-19 and Connections with Air Pollution Exposure: Evidence from the Netherlands [Preprint]. Epidemiology. https://doi.org/10.1101/2020.04.27.20081562

Archer, E., Petrie, B., Kasprzyk-Hordern, B., & Wolfaardt, G. M. (2017). The fate of pharmaceuticals and personal care products (PPCPs), endocrine disrupting contaminants (EDCs), metabolites and illicit drugs in a WWTW and environmental waters. Chemosphere, 174, 437–446. https://doi.org/10.1016/j.chemosphere.2017.01.101

Arpin-Pont, L., Bueno, M. J. M., Gomez, E., & Fenet, H. (2016). Occurrence of PPCPs in the marine environment: A review. Environmental Science and Pollution Research, 23(6), 4978–4991. https://doi.org/10.1007/s11356-014-3617-x

Asaro, R., & Lubarda, V. (2006). Mechanics of Solids and Materials (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511755514

Asghar, M. A., Zhu, Q., Sun, S., Peng, Y., & Shuai, Q. (2018). Suspect screening and target quantification of human pharmaceutical residues in the surface water of Wuhan, China, using UHPLC-Q-Orbitrap HRMS. Science of The Total Environment, 635, 828–837. https://doi.org/10.1016/j.scitotenv.2018.04.179

AstraZeneca. (2020). Environmental Risk Assessment Data—Metformin hydrochloride.

Bai, X., Lutz, A., Carroll, R., Keteles, K., Dahlin, K., Murphy, M., & Nguyen, D. (2018). Occurrence, distribution, and seasonality of emerging contaminants in urban watersheds. Chemosphere, 200, 133–142. https://doi.org/10.1016/j.chemosphere.2018.02.106

Balakrishnan, A., Sillanpää, M., Jacob, M. M., & Vo, D.-V. N. (2022). Metformin as an emerging concern in wastewater: Occurrence, analysis and treatment methods. Environmental Research, 213, 113613. https://doi.org/10.1016/j.envres.2022.113613

Balkhair, K. S. (2017). Modeling fecal bacteria transport and retention in agricultural and urban soils under saturated and unsaturated flow conditions. Water Research, 110, 313–320. https://doi.org/10.1016/j.watres.2016.12.023

Bandai, T., & Ghezzehei, T. A. (2021). Physics-Informed Neural Networks With Monotonicity Constraints for Richardson-Richards Equation: Estimation of Constitutive Relationships and Soil Water Flux Density From Volumetric Water Content Measurements. Water Resources Research, 57(2). https://doi.org/10.1029/2020WR027642

Bandai, T., & Ghezzehei, T. A. (2022). Forward and inverse modeling of water flow in unsaturated soils with discontinuous hydraulic conductivities using physics-informed neural networks with domain decomposition. Hydrology and Earth System Sciences, 26(16), 4469–4495. https://doi.org/10.5194/hess-26-4469-2022

Bao, R., & Zhang, A. (2020). Does lockdown reduce air pollution? Evidence from 44 cities in northern China. Science of The Total Environment, 731, 139052. https://doi.org/10.1016/j.scitotenv.2020.139052

Bartzen, B. A., Dufour, K. W., Clark, R. G., & Caswell, F. D. (2010). Trends in agricultural impact and recovery of wetlands in prairie Canada. Ecological Applications, 20(2), 525–538. https://doi.org/10.1890/08-1650.1

Barzegar, R., Moghaddam, A. A., Deo, R., Fijani, E., & Tziritis, E. (2018). Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Science of The Total Environment, 621, 697–712. https://doi.org/10.1016/j.scitotenv.2017.11.185

Bashir, M. F., Ma, B., Bilal, Komal, B., Bashir, M. A., Tan, D., & Bashir, M. (2020). Correlation between climate indicators and COVID-19 pandemic in New York, USA. Science of The Total Environment, 728, 138835. https://doi.org/10.1016/j.scitotenv.2020.138835

Bashir, M. F., Ma, B. J., Bilal, Komal, B., Bashir, M. A., Farooq, T. H., Iqbal, N., & Bashir, M. (2020). Correlation between environmental pollution indicators and COVID-19 pandemic: A brief study in Californian context. Environmental Research, 187, 109652. https://doi.org/10.1016/j.envres.2020.109652

Bates, S., Sesia, M., Sabatti, C., & Candès, E. (2020). Causal inference in genetic trio studies. Proceedings of the National Academy of Sciences, 117(39), 24117–24126. https://doi.org/10.1073/pnas.2007743117

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., & Syrgkanis, V. (2019). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. Microsoft. https://github.com/microsoft/EconML

Bear, J. (2013). Dynamics of Fluids in Porous Media. Dover Publications.

Bear, J., & Cheng, A. H.-D. (2010). Modeling Groundwater Flow and Contaminant Transport. Springer Netherlands. https://doi.org/10.1007/978-1-4020-6682-5

Bellman, R. E. (2015). Adaptive Control Processes: A Guided Tour. Princeton University Press.

Bengio, Y. (2012). Deep Learning of Representations for Unsupervised and Transfer Learning. Proceedings of ICML Workshop on Unsupervised and Transfer Learning, 27, 17–36.

Bertels, D., & Willems, P. (2023). Physics-informed machine learning method for modelling transport of a conservative pollutant in surface water systems. Journal of Hydrology, 619, 129354. https://doi.org/10.1016/j.jhydrol.2023.129354

Beven, K. (1996). Equifinality and Uncertainty in Geomorphological Modelling. The Scientific Nature of Geomorphology: Proceedings of the 27th Binghamton Symposium in Geomorphology, 27.

Beven, K. (2007). Towards integrated environmental models of everywhere: Uncertainty, data and modelling as a learning process. Hydrology and Earth System Sciences, 11(1), 460–467. https://doi.org/10.5194/hess-11-460-2007

Blyth, C. R. (1972). On Simpson's Paradox and the Sure-Thing Principle. Journal of the American Statistical Association, 67(338), 364–366. https://doi.org/10/gfw6js

Boehme, M. W. J., Buechele, G., Frankenhauser-Mannuss, J., Mueller, J., Lump, D., Boehm, B. O., & Rothenbacher, D. (2015). Prevalence, incidence and concomitant co-morbidities of type 2 diabetes mellitus in South Western Germany—A retrospective cohort and case control study in claims data of a large statutory health insurance. BMC Public Health, 15(1), 855. https://doi.org/10.1186/s12889-015-2188-1

Böhme, J., Martinez, N., Li, S., Lee, A., Marzuki, M., Tizazu, A. M., Ackart, D., Frenkel, J. H., Todd, A., Lachmandas, E., Lum, J., Shihui, F., Ng, T. P., Lee, B., Larbi, A., Netea, M. G., Basaraba, R., Van Crevel, R., Newell, E., … Singhal, A. (2020). Metformin enhances anti-mycobacterial responses by educating CD8+ T-cell immunometabolic circuits. Nature Communications, 11(1), 5225. https://doi.org/10.1038/s41467-020-19095-z

Bourouiba, L. (2020). Turbulent Gas Clouds and Respiratory Pathogen Emissions: Potential Implications for Reducing Transmission of COVID-19. JAMA, 323(18), 1837–1838. https://doi.org/10.1001/jama.2020.4756

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Briones, R. M., & Sarmah, A. K. (2018a). Detailed sorption characteristics of the anti-diabetic drug metformin and its transformation product guanylurea in agricultural soils. Science of The Total Environment, 630, 1258–1268. https://doi.org/10/gdqqc2

Briones, R. M., & Sarmah, A. K. (2018b). Insight into the sorption mechanism of metformin and its transformation product guanylurea in pastoral soils and model sorbents. Science of The Total Environment, 645, 1323–1333. https://doi.org/10/gfkpz4

Briones, R. M., & Sarmah, A. K. (2019). Sorption and mobility of metformin and guanylurea in soils as affected by biosolid amendment: Batch and column tests. Environmental Pollution, 244, 19–27. https://doi.org/10/gfsq7x

Briones, R. M., Sarmah, A. K., & Padhye, L. P. (2016). A global perspective on the use, occurrence, fate and effects of anti-diabetic drug metformin in natural and engineered ecosystems. Environmental Pollution, 219, 1007–1020. https://doi.org/10/f9hbqz

Briones, R. M., Zhuang, W.-Q., & Sarmah, A. K. (2018). Biodegradation of metformin and guanylurea by aerobic cultures enriched from sludge. Environmental Pollution, 243, 255–262. https://doi.org/10.1016/j.envpol.2018.08.075

Bromly, M., Hinz, C., & Aylmore, L. A. G. (2007). Relation of dispersivity to properties of homogeneous saturated repacked soil columns. European Journal of Soil Science, 58(1), 293–301. https://doi.org/10.1111/j.1365-2389.2006.00839.x

Brunetti, P., Baldessin, L., & Pagliacci, S. (2022). Prediabetes, undiagnosed diabetes and diabetes risk in Italy in 2017–2018: Results from the first National screening campaign in community pharmacies. Journal of Public Health, 44(3), 499–506. https://doi.org/10.1093/pubmed/fdab046

Brusseau, M. L., Jessup, R. E., & Rao, P. S. C. (1991). Nonequilibrium sorption of organic chemicals: Elucidation of rate-limiting processes. Environmental Science & Technology, 25(1), 134–142. https://doi.org/10.1021/es00013a015

Bruun-Rasmussen, N. E., Napolitano, G., Kofoed-Enevoldsen, A., Bojesen, S. E., Ellervik, C., Rasmussen, K., Jepsen, R., & Lynge, E. (2020). Burden of prediabetes, undiagnosed, and poorly or potentially sub-controlled diabetes:

Lolland-Falster health study. BMC Public Health, 20(1), 1711. https://doi.org/10.1186/s12889-020-09791-2

Burns, E. E., Carter, L. J., Kolpin, D. W., Thomas-Oates, J., & Boxall, A. B. A. (2018). Temporal and spatial variation in pharmaceutical concentrations in an urban river system. Water Research, 137, 72–85. https://doi.org/10.1016/j.watres.2018.02.066

Butcher, B., Huang, V. S., Robinson, C., Reffin, J., Sgaier, S. K., Charles, G., & Quadrianto, N. (2021). Causal Datasheet for Datasets: An Evaluation Guide for Real-World Data Analysis and Data Collection Design Using Bayesian Networks. Frontiers in Artificial Intelligence, 4. https://doi.org/10.3389/frai.2021.612551

Cai, S., Mao, Z., Wang, Z., Yin, M., & Karniadakis, G. E. (2021). Physics-informed neural networks (PINNs) for fluid mechanics: A review. Acta Mechanica Sinica, 37(12), 1727–1738. https://doi.org/10.1007/s10409-021-01148-1

Caldwell, D. J., D'Aco, V., Davidson, T., Kappler, K., Murray-Smith, R. J., Owen, S. F., Robinson, P. F., Simon-Hettich, B., Straub, J. O., & Tell, J. (2019). Environmental risk assessment of metformin and its transformation product guanylurea: II. Occurrence in surface waters of Europe and the United States and derivation of predicted no-effect concentrations. Chemosphere, 216, 855–865. https://doi.org/10.1016/j.chemosphere.2018.10.038

Callaghan, M., Schleussner, C.-F., Nath, S., Lejeune, Q., Knutson, T. R., Reichstein, M., Hansen, G., Theokritoff, E., Andrijevic, M., Brecha, R. J., Hegarty, M., Jones, C., Lee, K., Lucas, A., Van Maanen, N., Menke, I., Pfleiderer, P., Yesil, B., & Minx, J. C. (2021). Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. Nature Climate Change, 11(11), 966–972. https://doi.org/10.1038/s41558-021-01168-6

Cao, H., Xie, X., Shi, J., Jiang, G., & Wang, Y. (2022). Siamese Network-Based Transfer Learning Model to Predict Geogenic Contaminated Groundwaters. Environmental Science & Technology, 56(15), 11071–11079. https://doi.org/10.1021/acs.est.1c08682

Cardini, A., Pellegrino, E., & Ercoli, L. (2021). Predicted and Measured Concentration of Pharmaceuticals in Surface Water of Areas with Increasing Anthropic Pressure: A Case Study in the Coastal Area of Central Italy. Water, 13(20), 2807. https://doi.org/10.3390/w13202807

Carleton, T., Cornetet, J., Huybers, P., Meng, K. C., & Proctor, J. (2021). Global evidence for ultraviolet radiation decreasing COVID-19 growth rates. Proceedings of the National Academy of Sciences, 118(1). https://doi.org/10.1073/pnas.2012370118

Carmona, E., Andreu, V., & Picó, Y. (2017). Multi-residue determination of 47 organic compounds in water, soil, sediment and fish—Turia River as case study. Journal of Pharmaceutical and Biomedical Analysis, 146, 117–125. https://doi.org/10/gcgmmd

Carriger, J. F., Barron, M. G., & Newman, M. C. (2016). Bayesian Networks Improve Causal Environmental Assessments for Evidence-Based Policy. Environmental Science & Technology, 50(24), 13195–13205. https://doi.org/10.1021/acs.est.6b03220

Cerovečki, I., & Švajda, M. (2021). COVID-19 Pandemic Influence on Diabetes Management in Croatia. Frontiers in Clinical Diabetes and Healthcare, 2, 704807. https://doi.org/10.3389/fcdhc.2021.704807

Chambers, P. A., McGoldrick, D. J., Brua, R. B., Vis, C., Culp, J. M., & Benoy, G. A. (2012). Development of Environmental Thresholds for Nitrogen and Phosphorus in Streams. Journal of Environmental Quality, 41(1), 7–20. https://doi.org/10.2134/jeq2010.0273

Chan, H. M., Fediuk, K., Batal, M., Sadik, T., Tikhonov, C., Ing, A., & Barwin, L. (2021). The First Nations Food, Nutrition and Environment Study (2008–2018)—Rationale, design, methods and lessons learned. Canadian Journal of Public Health, 112(S1), 8–19. https://doi.org/10.17269/s41997-021-00480-0

Chan, J. F.-W., Yuan, S., Kok, K.-H., To, K. K.-W., Chu, H., Yang, J., Xing, F., Liu, J., Yip, C. C.-Y., Poon, R. W.-S., Tsoi, H.-W., Lo, S. K.-F., Chan, K.-H., Poon, V. K.-M., Chan, W.-M., Ip, J. D., Cai, J.-P., Cheng, V. C.-C., Chen, H., … Yuen, K.-Y. (2020). A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: A study of a family cluster. The Lancet, 395(10223), 514–523. https://doi.org/10.1016/S0140-6736(20)30154-9

Chan, L., Batal, M., Sadik, T., Tikhonov, C., Schwartz, H., Fediuk, K., Ing, A., Marushka, L., Lindhorst, K., Barwin, L., Berti, P., Singh, K., & Receveur, O. (2019). FNFNES Final Report for Eight Assembly of First Nations Regions: Draft Comprehensive Technical Report. Assembly of First Nations, University of Ottawa, Université de Montréal.

Chattopadhyay, A., Manupriya, P., Sarkar, A., & Balasubramanian, V. N. (2019). Neural Network Attributions: A Causal Perspective. Proceedings of the 36th International Conference on Machine Learning, 97. https://proceedings.mlr.press/v97/chattopadhyay19a.html

Chaves-Barquero, L. G., Luong, K. H., Mundy, C. J., Knapp, C. W., Hanson, M. L., & Wong, C. S. (2016). The release of wastewater contaminants in the Arctic: A case study from Cambridge Bay, Nunavut, Canada. Environmental Pollution, 218, 542–550. https://doi.org/10.1016/j.envpol.2016.07.036

Chen, S., Liang, Z., Webster, R., Zhang, G., Zhou, Y., Teng, H., Hu, B., Arrouays, D., & Shi, Z. (2019). A high-resolution map of soil pH in China made by hybrid modelling of sparse soil data and environmental covariates and its implications for pollution. Science of The Total Environment, 655, 273–283. https://doi.org/10.1016/j.scitotenv.2018.11.230

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Xu, H., Jiang, P., Yu, S., Lin, G., Bychkov, I., Hmelnov, A., Ruzhnikov, G., Zhu, N., & Liu, Z. (2021). A transfer Learning-Based LSTM strategy for imputing Large-Scale consecutive missing data and its application in a water quality prediction system. Journal of Hydrology, 602, 126573. https://doi.org/10.1016/j.jhydrol.2021.126573

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2017). Double/Debiased Machine Learning for Treatment and Causal Parameters. arXiv:1608.00060 [Econ, Stat]. http://arxiv.org/abs/1608.00060

Chia, P. Y., Coleman, K. K., Tan, Y. K., Ong, S. W. X., Gum, M., Lau, S. K., Lim, X. F., Lim, A. S., Sutjipto, S., Lee, P. H., Son, T. T., Young, B. E., Milton, D. K., Gray, G. C., Schuster, S., Barkham, T., De, P. P., Vasoo, S., Chan, M., … Marimuthu, K. (2020). Detection of air and surface contamination by SARS-CoV-2 in hospital rooms of infected patients. Nature Communications, 11(1), 2800. https://doi.org/10.1038/s41467-020-16670-2

Choi, S., Yoom, H., Son, H., Seo, C., Kim, K., Lee, Y., & Kim, Y. M. (2022). Removal efficiency of organic micropollutants in successive wastewater treatment steps in a full-scale wastewater treatment plant: Bench-scale application of tertiary treatment processes to improve removal of organic

micropollutants persisting after secondary treatment. Chemosphere, 288, 132629. https://doi.org/10.1016/j.chemosphere.2021.132629

Choi, Y., Lee, J.-H., Kim, K., Mun, H., Park, N., & Jeon, J. (2021). Identification, quantification, and prioritization of new emerging pollutants in domestic and industrial effluents, Korea: Application of LC-HRMS based suspect and non-target screening. Journal of Hazardous Materials, 402, 123706. https://doi.org/10.1016/j.jhazmat.2020.123706

Coccia, M. (2020a). Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. Science of The Total Environment, 729, 138474. https://doi.org/10.1016/j.scitotenv.2020.138474

Coccia, M. (2020b). How (Un)sustainable Environments Are Related to the Diffusion of COVID-19: The Relation between Coronavirus Disease 2019, Air Pollution, Wind Resource and Energy. Sustainability, 12(22), Article 22. https://doi.org/10.3390/su12229709

Coccia, M. (2021a). Effects of the spread of COVID-19 on public health of polluted cities: Results of the first wave for explaining the dejà vu in the second wave of COVID-19 pandemic and epidemics of future vital agents. Environmental Science and Pollution Research, 28(15), 19147–19154. https://doi.org/10.1007/s11356-020-11662-7

Coccia, M. (2021b). How do low wind speeds and high levels of air pollution support the spread of COVID-19? Atmospheric Pollution Research, 12(1), 437–445. https://doi.org/10.1016/j.apr.2020.10.002

Coccia, M. (2021c). The effects of atmospheric stability with low wind speed and of air pollution on the accelerated transmission dynamics of COVID-19. International Journal of Environmental Studies, 78(1), 1–27. https://doi.org/10.1080/00207233.2020.1802937

Coccia, M. (2021d). The impact of first and second wave of the COVID-19 pandemic in society: Comparative analysis to support control measures to cope with negative effects of future infectious diseases. Environmental Research, 197, 111099. https://doi.org/10.1016/j.envres.2021.111099

Coccia, M. (2021e). The relation between length of lockdown, numbers of infected people and deaths of Covid-19, and economic growth of countries: Lessons learned to cope with future pandemics similar to Covid-19 and to constrain the deterioration of economic system. Science of The Total Environment, 775, 145801. https://doi.org/10.1016/j.scitotenv.2021.145801

Cole, M. A., Elliott, R. J. R., & Liu, B. (2020). The Impact of the Wuhan Covid-19 Lockdown on Air Pollution and Health: A Machine Learning and Augmented Synthetic Control Approach. Environmental and Resource Economics, 76(4), 553–580. https://doi.org/10.1007/s10640-020-00483-4

Collenteur, R., Vremec, M., & Brunetti, G. (2020). Interfacing FORTAN Code with Python: An example for the Hydrus-1D model (EGU2020-15377). EGU2020. Copernicus Meetings. https://doi.org/10.5194/egusphere-egu2020-15377

Damette, O., & Goutte, S. (2020). Weather, pollution and Covid-19 spread: A time series and Wavelet reassessment (pp. 1–22). https://halshs.archives-ouvertes.fr/halshs-02629139

Davis, S. J., Liu, Z., Deng, Z., Zhu, B., Ke, P., Sun, T., Guo, R., Hong, C., Zheng, B., Wang, Y., Boucher, O., Gentine, P., & Ciais, P. (2022). Emissions rebound from the COVID-19 pandemic. Nature Climate Change, 12(5), 412–414. https://doi.org/10.1038/s41558-022-01332-6

De Jesus Gaffney, V., Cardoso, V. V., Cardoso, E., Teixeira, A. P., Martins, J., Benoliel, M. J., & Almeida, C. M. M. (2017). Occurrence and behaviour of pharmaceutical compounds in a Portuguese wastewater treatment plant: Removal efficiency through conventional treatment processes. Environmental Science and Pollution Research, 24(17), 14717–14734. https://doi.org/10.1007/s11356-017-9012-7

De Mestral, C., Stringhini, S., Guessous, I., & Jornayvaz, F. R. (2020). Thirteen-year trends in the prevalence of diabetes in an urban region of Switzerland: A population-based study. Diabetic Medicine, 37(8), 1374–1378. https://doi.org/10.1111/dme.14206

Deblonde, T., Cossu-Leguille, C., & Hartemann, P. (2011). Emerging pollutants in wastewater: A review of the literature. International Journal of Hygiene and Environmental Health, 214(6), 442–448. https://doi.org/10/chq9ds

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. Journal of the American Statistical Association, 94(448), 1053–1062. https://doi.org/10.1080/01621459.1999.10473858

Delnevo, G., Mirri, S., & Roccetti, M. (2020). Particulate Matter and COVID-19 Disease Diffusion in Emilia-Romagna (Italy). Already a Cold Case? Computation, 8(2), Article 2. https://doi.org/10.3390/computation8020059

Diamond, J. M., Latimer, H. A., Munkittrick, K. R., Thornton, K. W., Bartell, S. M., & Kidd, K. A. (2011). Prioritizing contaminants of emerging concern for ecological screening assessments. Environmental Toxicology and Chemistry, 30(11), 2385–2394. https://doi.org/10/bmb3ph

Diao, Y., Kodera, S., Anzai, D., Gomez-Tames, J., Rashed, E. A., & Hirata, A. (2021). Influence of population density, temperature, and absolute humidity on spread and decay durations of COVID-19: A comparative study of scenarios in China, England, Germany, and Japan. One Health, 12, 100203. https://doi.org/10.1016/j.onehlt.2020.100203

Domenico, P. A., & Robbins, G. A. (1984). A dispersion scale effect in model calibrations and field tracer experiments. Journal of Hydrology, 70(1–4), 123–132. https://doi.org/10.1016/0022-1694(84)90117-3

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases, 20(5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1

Dowling, R. J. O., Niraula, S., Stambolic, V., & Goodwin, P. J. (2012). Metformin in cancer: Translational challenges. Journal of Molecular Endocrinology, 48(3), R31–R43. https://doi.org/10/ggf22k

Drever, C. R., Cook-Patton, S. C., Akhter, F., Badiou, P. H., Chmura, G. L., Davidson, S. J., Desjardins, R. L., Dyk, A., Fargione, J. E., Fellows, M., Filewod, B., Hessing-Lewis, M., Jayasundara, S., Keeton, W. S., Kroeger, T., Lark, T. J., Le, E., Leavitt, S. M., LeClerc, M.-E., … Kurz, W. A. (2021). Natural climate solutions for Canada. Science Advances, 7(23), eabd6034. https://doi.org/10.1126/sciadv.abd6034

Drzewoski, J., & Hanefeld, M. (2021). The Current and Potential Therapeutic Use of Metformin—The Good Old Drug. Pharmaceuticals, 14(2), 122. https://doi.org/10.3390/ph14020122

Dulio, V., van Bavel, B., Brorström-Lundén, E., Harmsen, J., Hollender, J., Schlabach, M., Slobodnik, J., Thomas, K., & Koschorreck, J. (2018). Emerging pollutants in the EU: 10 years of NORMAN in support of environmental policies and regulations. Environmental Sciences Europe, 30(1), 5. https://doi.org/10.1186/s12302-018-0135-3

Ebert-Uphoff, I., & Deng, Y. (2017). Causal discovery in the geosciences—Using synthetic data to learn how to interpret results. Computers & Geosciences, 99, 50–60. https://doi.org/10.1016/j.cageo.2016.10.008

Ehalt Macedo, H., Lehner, B., Nicell, J., Grill, G., Li, J., Limtong, A., & Shakya, R. (2021). Global distribution of wastewater treatment plants and their released effluents into rivers and streams [Preprint]. Hydrology and Soil Science – Hydrology. https://doi.org/10.5194/essd-2021-214

EL-Arabey, A. A., & Abdalla, M. (2020). Metformin and COVID-19: A novel deal of an old drug. Journal of Medical Virology, 92(11), 2293–2294. https://doi.org/10.1002/jmv.25958

Elizalde-Velázquez, G. A., & Gómez-Oliván, L. M. (2020). Occurrence, toxic effects and removal of metformin in the aquatic environments in the world: Recent trends and perspectives. Science of The Total Environment, 702, 134924. https://doi.org/10.1016/j.scitotenv.2019.134924

Estrada-Arriaga, E. B., Cortés-Muñoz, J. E., González-Herrera, A., Calderón-Mólgora, C. G., de Lourdes Rivera-Huerta, Ma., Ramírez-Camperos, E., Montellano-Palacios, L., Gelover-Santiago, S. L., Pérez-Castrejón, S., Cardoso-Vigueros, L., Martín-Domínguez, A., & García-Sánchez, L. (2016). Assessment of full-scale biological nutrient removal systems upgraded with physico-chemical processes for the removal of emerging pollutants present in wastewaters from Mexico. Science of The Total Environment, 571, 1172–1182. https://doi.org/10/f848t2

Fahey, D. W., Hübler, G., Parrish, D. D., Williams, E. J., Norton, R. B., Ridley, B. A., Singh, H. B., Liu, S. C., & Fehsenfeld, F. C. (1986). Reactive nitrogen species in the troposphere: Measurements of NO, NO2, HNO3, particulate nitrate, peroxyacetyl nitrate (PAN), O3, and total reactive odd nitrogen (NO y) at Niwot Ridge, Colorado. Journal of Geophysical Research: Atmospheres, 91(D9), 9781–9793. https://doi.org/10.1029/JD091iD09p09781

Fan, Y. R. (2022). Bivariate hydrologic risk analysis for the Xiangxi River in Three Gorges Reservoir Area, China. Environmental Systems Research, 11(1), 18. https://doi.org/10.1186/s40068-022-00264-6

Fang, M., Wang, D., Coresh, J., & Selvin, E. (2022). Undiagnosed Diabetes in U.S. Adults: Prevalence and Trends. Diabetes Care, 45(9), 1994–2002. https://doi.org/10.2337/dc22-0242

Fears, A. C., Klimstra, W. B., Duprex, P., Hartman, A., Weaver, S. C., Plante, K. S., Mirchandani, D., Plante, J. A., Aguilar, P. V., Fernández, D., Nalca, A., Totura, A., Dyer, D., Kearney, B., Lackemeyer, M., Bohannon, J. K., Johnson, R., Garry, R. F., Reed, D. S., & Roy, C. J. (2020). Persistence of Severe Acute

Respiratory Syndrome Coronavirus 2 in Aerosol Suspensions. Emerging Infectious Diseases, 26(9), 2168–2171. https://doi.org/10.3201/eid2609.201806

Feizi, F., Sarmah, A. K., & Rangsivek, R. (2021). Adsorption of pharmaceuticals in a fixed-bed column using tyre-based activated carbon: Experimental investigations and numerical modelling. Journal of Hazardous Materials, 417, 126010. https://doi.org/10.1016/j.jhazmat.2021.126010

Fleming, S. W., Watson, J. R., Ellenson, A., Cannon, A. J., & Vesselinov, V. C. (2021). Machine learning in Earth and environmental science requires education and research policy reforms. Nature Geoscience, 14(12), 878–880. https://doi.org/10.1038/s41561-021-00865-3

Fong, I. H., Li, T., Fong, S., Wong, R. K., & Tallón-Ballesteros, A. J. (2020). Predicting concentration levels of air pollutants by transfer learning and recurrent neural network. Knowledge-Based Systems, 192, 105622. https://doi.org/10.1016/j.knosys.2020.105622

Forster, P. M., Forster, H. I., Evans, M. J., Gidden, M. J., Jones, C. D., Keller, C. A., Lamboll, R. D., Quéré, C. L., Rogelj, J., Rosen, D., Schleussner, C.-F., Richardson, T. B., Smith, C. J., & Turnock, S. T. (2020). Current and future global climate impacts resulting from COVID-19. Nature Climate Change, 10(10), 913–919. https://doi.org/10/gg7s7w

Foster, S. S. D., & Chilton, P. J. (2003). Groundwater: The processes and global significance of aquifer degradation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358(1440), 1957–1972. https://doi.org/10.1098/rstb.2003.1380

Freeze, R. A., & Cherry, J. A. (1979). Groundwater. Prentice-Hall.

Gao, G., Zhan, H., Feng, S., Fu, B., Ma, Y., & Huang, G. (2010). A new mobile-immobile model for reactive solute transport with scale-dependent dispersion: MOBILE-IMMOBILE MODEL. Water Resources Research, 46(8). https://doi.org/10.1029/2009WR008707

Gao, W., Tie, X., Xu, J., Huang, R., Mao, X., Zhou, G., & Chang, L. (2017). Long-term trend of O3 in a mega City (Shanghai), China: Characteristics, causes, and interactions with precursors. Science of The Total Environment, 603–604, 425–433. https://doi.org/10.1016/j.scitotenv.2017.06.099

Geissen, V., Mol, H., Klumpp, E., Umlauf, G., Nadal, M., Van Der Ploeg, M., Van De Zee, S. E. A. T. M., & Ritsema, C. J. (2015). Emerging pollutants in the environment: A challenge for water resource management. International Soil

and Water Conservation Research, 3(1), 57–65. https://doi.org/10.1016/j.iswcr.2015.03.002

Gelhar, L. W., & Axness, C. L. (1983). Three-dimensional stochastic analysis of macrodispersion in aquifers. Water Resources Research, 19(1), 161–180. https://doi.org/10.1029/WR019i001p00161

Gelhar, L. W., Welty, C., & Rehfeldt, K. R. (1992). A critical review of data on field-scale dispersion in aquifers. Water Resources Research, 28(7), 1955–1974. https://doi.org/10.1029/92WR00607

Ghorbani, A., Sadeghi, M., & Jones, S. B. (2021). Towards new soil water flow equations using physics-constrained machine learning. Vadose Zone Journal, 20(4). https://doi.org/10.1002/vzj2.20136

Ghoshdastidar, A. J., Fox, S., & Tong, A. Z. (2015). The presence of the top prescribed pharmaceuticals in treated sewage effluents and receiving waters in Southwest Nova Scotia, Canada. Environmental Science and Pollution Research, 22(1), 689–700. https://doi.org/10.1007/s11356-014-3400-z

Gibert, K., Horsburgh, J. S., Athanasiadis, I. N., & Holmes, G. (2018). Environmental Data Science. Environmental Modelling & Software, 106, 4–12. https://doi.org/10.1016/j.envsoft.2018.04.005

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. Frontiers in Genetics, 10. https://doi.org/10.3389/fgene.2019.00524

Godoy, A. A., Domingues, I., Arsénia Nogueira, A. J., & Kummrow, F. (2018). Ecotoxicological effects, water quality standards and risk assessment for the anti-diabetic metformin. Environmental Pollution, 243, 534–542. https://doi.org/10.1016/j.envpol.2018.09.031

Golovko, O., Örn, S., Sörengård, M., Frieberg, K., Nassazzi, W., Lai, F. Y., & Ahrens, L. (2021). Occurrence and removal of chemicals of emerging concern in wastewater treatment plants and their impact on receiving water systems. Science of The Total Environment, 754, 142122. https://doi.org/10.1016/j.scitotenv.2020.142122

González-Gaya, B., Lopez-Herguedas, N., Santamaria, A., Mijangos, F., Etxebarria, N., Olivares, M., Prieto, A., & Zuloaga, O. (2021). Suspect screening workflow comparison for the analysis of organic xenobiotics in environmental water samples. Chemosphere, 274, 129964. https://doi.org/10.1016/j.chemosphere.2021.129964

Griebler, C., & Lueders, T. (2009). Microbial biodiversity in groundwater ecosystems. Freshwater Biology, 54(4), 649–677. https://doi.org/10.1111/j.1365-2427.2008.02013.x

Groffman, P. M., Howard, G., Gold, A. J., & Nelson, W. M. (1996). Microbial Nitrate Processing in Shallow Groundwater in a Riparian Forest. Journal of Environmental Quality, 25(6), 1309–1316. https://doi.org/10.2134/jeq1996.00472425002500060020x

Gunnarsdóttir, R., Jenssen, P. D., Erland Jensen, P., Villumsen, A., & Kallenborn, R. (2013). A review of wastewater handling in the Arctic with special reference to pharmaceuticals and personal care products (PPCPs) and microbial pollution. Ecological Engineering, 50, 76–85. https://doi.org/10.1016/j.ecoleng.2012.04.025

Guo, Z.-D., Wang, Z.-Y., Zhang, S.-F., Li, X., Li, L., Li, C., Cui, Y., Fu, R.-B., Dong, Y.-Z., Chi, X.-Y., Zhang, M.-Y., Liu, K., Cao, C., Liu, B., Zhang, K., Gao, Y.-W., Lu, B., & Chen, W. (2020). Aerosol and Surface Distribution of Severe Acute Respiratory Syndrome Coronavirus 2 in Hospital Wards, Wuhan, China, 2020. Emerging Infectious Diseases, 26(7), 1586–1591. https://doi.org/10.3201/eid2607.200885

Hao, P., Di, L., Zhang, C., & Guo, L. (2020). Transfer Learning for Crop classification with Cropland Data Layer data (CDL) as training samples. Science of The Total Environment, 733, 138869. https://doi.org/10.1016/j.scitotenv.2020.138869

Haque, S. E., & Rahman, M. (2020). Association between temperature, humidity, and COVID-19 outbreaks in Bangladesh. Environmental Science & Policy, 114, 253–255. https://doi.org/10.1016/j.envsci.2020.08.012

He, G., Pan, Y., & Tanaka, T. (2020). The short-term impacts of COVID-19 lockdown on urban air pollution in China. Nature Sustainability. https://doi.org/10.1038/s41893-020-0581-y

He, S., Wu, J., Wang, D., & He, X. (2022). Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. Chemosphere, 290, 133388. https://doi.org/10.1016/j.chemosphere.2021.133388

He, Y., Jin, H., Gao, H., Zhang, G., & Ju, F. (2022). Prevalence, production, and ecotoxicity of chlorination-derived metformin byproducts in Chinese urban water systems. Science of The Total Environment, 816, 151665. https://doi.org/10.1016/j.scitotenv.2021.151665

He, Y., Zhang, Y., & Ju, F. (2022). Metformin Contamination in Global Waters: Biotic and Abiotic Transformation, Byproduct Generation and Toxicity, and Evaluation as a Pharmaceutical Indicator. Environmental Science & Technology, 56(19), 13528–13545. https://doi.org/10.1021/acs.est.2c02495

Heberling, J. M., Miller, J. T., Noesgaard, D., Weingart, S. B., & Schigel, D. (2021). Data integration enables global biodiversity synthesis. Proceedings of the National Academy of Sciences, 118(6). https://doi.org/10.1073/pnas.2018093118

Heckman, J. J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. In Journal of Economic and Social Measurement (Vol. 5, pp. 475–492). NBER.

Heckman, J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. Econometrica, 46(4), 931–959. https://doi.org/10.2307/1909757

Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association, 81(396), 945–960. https://doi.org/10.1080/01621459.1986.10478354

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X., Cheng, Z., Yu, T., Xia, J., Wei, Y., Wu, W., Xie, X., Yin, W., Li, H., Liu, M., … Cao, B. (2020). Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet, 395(10223), 497–506. https://doi.org/10.1016/S0140-6736(20)30183-5

Huang, Q. (2010). An Integrated MM5-CAMx Modeling Approach for Assessing PM10 Contribution from Different Sources in Beijing, China. Journal of Environmental Informatics. https://doi.org/10.3808/jei.201000166

Huang, Y., & Valtorta, M. (2012). Pearl's Calculus of Intervention Is Complete. arXiv:1206.6831 [Cs]. http://arxiv.org/abs/1206.6831

Hudson, N. (1993). Field Measurement of Soil Erosion and Runoff (Vol. 68). Food and Agriculture Organization of the United Nations.

ICPDR. (2020). The Fourth Joint Danube Survey Scientific Report. ICPDR – International Commission for the Protection of the Danube River.

Imbens, G. W., & Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press. www.cambridge.org/9780521885881

Inarmal, N., & Moodley, B. (2023). Selected pharmaceutical analysis in a wastewater treatment plant during COVID-19 infection waves in South Africa. Environmental Science: Water Research & Technology, 9(6), 1566–1576. https://doi.org/10.1039/D3EW00059A

Ireson, A. M., Barr, A. G., Johnstone, J. F., Mamet, S. D., Van Der Kamp, G., Whitfield, C. J., Michel, N. L., North, R. L., Westbrook, C. J., DeBeer, C., Chun, K. P., Nazemi, A., & Sagin, J. (2015). The changing water cycle: The Boreal Plains ecozone of Western Canada. WIREs Water, 2(5), 505–521. https://doi.org/10.1002/wat2.1098

Islam, N., Bukhari, Q., Jameel, Y., Shabnam, S., Erzurumluoglu, A. M., Siddique, M. A., Massaro, J. M., & D'Agostino, R. B. (2021). COVID-19 and climatic factors: A global analysis. Environmental Research, 193, 110355. https://doi.org/10.1016/j.envres.2020.110355

Jacob, S., Köhler, H.-R., Tisler, S., Zwiener, C., & Triebskorn, R. (2019). Impact of the Antidiabetic Drug Metformin and Its Transformation Product Guanylurea on the Health of the Big Ramshorn Snail (Planorbarius corneus). Frontiers in Environmental Science, 7, 45. https://doi.org/10.3389/fenvs.2019.00045

Jeong, S., Fischer, M. L., Breunig, H., Marklein, A. R., Hopkins, F. M., & Biraud, S. C. (2022). Artificial Intelligence Approach for Estimating Dairy Methane Emissions. Environmental Science & Technology, 56(8), 4849–4858. https://doi.org/10.1021/acs.est.1c08802

Jing, L., Chen, B., Zhang, B., & Ye, X. (2018). Modeling marine oily wastewater treatment by a probabilistic agent-based approach. Marine Pollution Bulletin, 127, 217–224. https://doi.org/10.1016/j.marpolbul.2017.12.004

Ju, F., Beck, K., Yin, X., Maccagnan, A., McArdell, C. S., Singer, H. P., Johnson, D. R., Zhang, T., & Bürgmann, H. (2019). Wastewater treatment plant resistomes are shaped by bacterial composition, genetic exchange, and upregulated expression in the effluent microbiomes. The ISME Journal, 13(2), 346–360. https://doi.org/10.1038/s41396-018-0277-8

Kalainathan, D., & Goudet, O. (2019). Causal Discovery Toolbox: Uncover causal relationships in Python. arXiv:1903.02278 [Stat]. http://arxiv.org/abs/1903.02278

Kalainathan, D., Goudet, O., Guyon, I., Lopez-Paz, D., & Sebag, M. (2020). Structural Agnostic Modeling: Adversarial Learning of Causal Graphs. arXiv:1803.04929 [Stat]. http://arxiv.org/abs/1803.04929

Kancheti, S. S., Reddy, A. G., Balasubramanian, V. N., & Sharma, A. (2022). Matching Learned Causal Effects of Neural Networks with Domain Priors (arXiv:2111.12490). arXiv. http://arxiv.org/abs/2111.12490

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. Nature Reviews Physics, 3(6), 422–440. https://doi.org/10.1038/s42254-021-00314-5

Kassani, P. H., Lu, F., Le Guen, Y., Belloy, M. E., & He, Z. (2022). Deep neural networks with controlled variable selection for the identification of putative causal genetic variants. Nature Machine Intelligence, 4(9), 761–771. https://doi.org/10.1038/s42256-022-00525-0

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems 30 (NIPS 2017).

Kirby, T. (2021). New variant of SARS-CoV-2 in UK causes surge of COVID-19. The Lancet Respiratory Medicine, 9(2), e20–e21. https://doi.org/10.1016/S2213-2600(21)00005-9

Koroša, A., Brenčič, M., & Mali, N. (2020). Estimating the transport parameters of propyphenazone, caffeine and carbamazepine by means of a tracer experiment in a coarse-gravel unsaturated zone. Water Research, 175, 115680. https://doi.org/10.1016/j.watres.2020.115680

Kot-Wasik, A., Jakimska, A., & Śliwka-Kaszyńska, M. (2016). Occurrence and seasonal variations of 25 pharmaceutical residues in wastewater and drinking water treatment plants. Environmental Monitoring and Assessment, 188(12), 661. https://doi.org/10.1007/s10661-016-5637-0

Kulkarni, H., Khandait, H., Narlawar, U. W., Rathod, P., & Mamtani, M. (2020). Independent association of meteorological characteristics with initial spread of Covid-19 in India. Science of The Total Environment, 142801. https://doi.org/10/ghqh38

Kumar, C. P. (2012). Groundwater Modelling Software – Capabilities and Limitations. IOSR Journal of Environmental Science, Toxicology and Food Technology, 1(2), 46–57. https://doi.org/10/gfkr68

Kutter, J. S., de Meulder, D., Bestebroer, T. M., Lexmond, P., Mulders, A., Richard, M., Fouchier, R. A. M., & Herfst, S. (2021). SARS-CoV and SARS-CoV-2 are transmitted through the air between ferrets over more than one meter

distance. Nature Communications, 12(1), Article 1. https://doi.org/10.1038/s41467-021-21918-6

La Farré, M., Pérez, S., Kantiani, L., & Barceló, D. (2008). Fate and toxicity of emerging pollutants, their metabolites and transformation products in the aquatic environment. TrAC Trends in Analytical Chemistry, 27(11), 991–1007. https://doi.org/10.1016/j.trac.2008.09.010

Laranjo, L., Rodrigues, D., Pereira, A. M., Ribeiro, R. T., & Boavida, J. M. (2016). Use of Electronic Health Records and Geographic Information Systems in Public Health Surveillance of Type 2 Diabetes: A Feasibility Study. JMIR Public Health and Surveillance, 2(1), e12. https://doi.org/10.2196/publichealth.4319

Lertxundi, U., Domingo-Echaburu, S., Barros, S., Santos, M. M., Neuparth, T., Quintana, J. B., Rodil, R., Montes, R., & Orive, G. (2023). Is the Environmental Risk of Metformin Underestimated? Environmental Science & Technology, 57(23), 8463–8466. https://doi.org/10.1021/acs.est.3c02468

Lesser, L. E., Mora, A., Moreau, C., Mahlknecht, J., Hernández-Antonio, A., Ramírez, A. I., & Barrios-Piña, H. (2018). Survey of 218 organic contaminants in groundwater derived from the world's largest untreated wastewater irrigation system: Mezquital Valley, Mexico. Chemosphere, 198, 510–521. https://doi.org/10.1016/j.chemosphere.2018.01.154

Li, J., Guo, K., Cao, Y., Wang, S., Song, Y., & Zhang, H. (2021). Enhance in mobility of oxytetracycline in a sandy loamy soil caused by the presence of microplastics. Environmental Pollution, 269, 116151. https://doi.org/10.1016/j.envpol.2020.116151

Li, P., Wu, H. J., & Chen, B. (2013). RSW-MCFP: A Resource-Oriented Solid Waste Management System for a Mixed Rural-Urban Area through Monte Carlo Simulation-Based Fuzzy Programming. Mathematical Problems in Engineering, 2013, 1–15. https://doi.org/10/gbdmc6

Lin, W., Yan, Y., Ping, S., Li, P., Li, D., Hu, J., Liu, W., Wen, X., & Ren, Y. (2021). Metformin-Induced Epigenetic Toxicity in Zebrafish: Experimental and Molecular Dynamics Simulation Studies. Environmental Science & Technology, 55(3), 1672–1681. https://doi.org/10.1021/acs.est.0c06052

Littlejohn, C., Renaud, J. B., Sabourin, L., Lapen, D. R., Pappas, J. J., Tuteja, B., Hughes, D., Ussery, E., Yeung, K. K. -C., & Sumarah, M. W. (2023). Environmental Concentrations of the Type 2 Diabetes Medication Metformin and Its Transformation Product Guanylurea in Surface Water and Sediment in

Ontario and Quebec, Canada. Environmental Toxicology and Chemistry, 42(8), 1709–1720. https://doi.org/10.1002/etc.5684

Liu, F., Wang, M., & Zheng, M. (2021). Effects of COVID-19 lockdown on global air quality and health. Science of The Total Environment, 755, 142533. https://doi.org/10.1016/j.scitotenv.2020.142533

LIU, H., ZHANG, M., & HAN, X. (2020). A review of surface ozone source apportionment in China. Atmospheric and Oceanic Science Letters, 0(0), 1–15. https://doi.org/10.1080/16742834.2020.1768025

Liu, J., Lipsitt, J., Jerrett, M., & Zhu, Y. (2020). Decreases in Near-Road NO and NO 2 Concentrations during the COVID-19 Pandemic in California. Environmental Science & Technology Letters, acs.estlett.0c00815. https://doi.org/10.1021/acs.estlett.0c00815

Liu, J.-Y., Woodward, R. T., & Zhang, Y.-J. (2021). Has Carbon Emissions Trading Reduced PM2.5 in China? Environmental Science & Technology, 55(10), 6631–6643. https://doi.org/10.1021/acs.est.1c00248

Lopez, B., Ollivier, P., Togola, A., Baran, N., & Ghestem, J.-P. (2015). Screening of French groundwater for regulated and emerging contaminants. Science of The Total Environment, 518–519, 562–573. https://doi.org/10.1016/j.scitotenv.2015.01.110

Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., & Welling, M. (2017). Causal Effect Inference with Deep Latent-Variable Models. arXiv:1705.08821 [Cs, Stat]. http://arxiv.org/abs/1705.08821

Lovrić, M., Pavlović, K., Vuković, M., Grange, S. K., Haberl, M., & Kern, R. (2020). Understanding the true effects of the COVID-19 lockdown on air pollution by means of machine learning. Environmental Pollution, 115900. https://doi.org/10/ghrsng

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), Article 1. https://doi.org/10.1038/s42256-019-0138-9

Luo, Y., Peng, J., & Ma, J. (2020). When causal inference meets deep learning. Nature Machine Intelligence, 2(8), 426–427. https://doi.org/10.1038/s42256-020-0218-x

Ma, W., Yuan, Z., Lau, A. K. H., Wang, L., Liao, C., & Zhang, Y. (2022). Optimized neural network for daily-scale ozone prediction based on transfer

learning. Science of The Total Environment, 827, 154279. https://doi.org/10.1016/j.scitotenv.2022.154279

Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., Yan, J., Niu, J., Zhou, J., & Luo, B. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. Science of The Total Environment, 724, 138226. https://doi.org/10.1016/j.scitotenv.2020.138226

MacLaren, R. D., Wisniewski, K., & MacLaren, C. (2018). Environmental concentrations of metformin exposure affect aggressive behavior in the Siamese fighting fish, Betta splendens. PLOS ONE, 13(5), e0197259. https://doi.org/10/gdjj56

MacQuarrie, K. T. B., Sudicky, E. A., & Robertson, W. D. (2001). Numerical simulation of a fine-grained denitrification layer for removing septic system nitrate from shallow groundwater. Journal of Contaminant Hydrology, 52(1–4), 29–55. https://doi.org/10.1016/S0169-7722(01)00152-8

Magliano, D., & Boyko, E. J. (2021). IDF diabetes atlas (10th edition). International Diabetes Federation.

Maraqa, M. A., Zhao, X., Lee, J., Allan, F., & Voice, T. C. (2011). Comparison of nonideal sorption formulations in modeling the transport of phthalate esters through packed soil columns. Journal of Contaminant Hydrology, 125(1–4), 57–69. https://doi.org/10.1016/j.jconhyd.2011.05.001

Markiewicz, M., Jungnickel, C., Stolte, S., Białk-Bielińska, A., Kumirska, J., & Mrozik, W. (2017). Ultimate biodegradability and ecotoxicity of orally administered antidiabetic drugs. Journal of Hazardous Materials, 333, 154–161. https://doi.org/10/f97ntm

Marshall, I. B., Smith, C. A. S., & Selby, C. J. (1996). A National Framework for Monitoring and Reporting on Environmental Sustainability in Canada. In R. A. Sims, I. G. W. Corns, & K. Klinka (Eds.), Global to Local: Ecological Land Classification (pp. 25–38). Springer Netherlands. https://doi.org/10.1007/978-94-009-1653-1_4

Mastakouri, A. A., & Schölkopf, B. (2020). Causal analysis of Covid-19 spread in Germany. arXiv:2007.11896 [Stat]. http://arxiv.org/abs/2007.11896

Mele, M., & Magazzino, C. (2020). Pollution, economic growth, and COVID-19 deaths in India: A machine learning evidence. Environmental Science and Pollution Research. https://doi.org/10.1007/s11356-020-10689-0

Melvin, S. D., Habener, L. J., Leusch, F. D. L., & Carroll, A. R. (2017). 1 H NMR-based metabolomics reveals sub-lethal toxicity of a mixture of diabetic and lipid-regulating pharmaceuticals on amphibian larvae. Aquatic Toxicology, 184, 123–132. https://doi.org/10.1016/j.aquatox.2017.01.012

Menéndez Torre, E. L., Blanco, J. A., Barreiro, S. C., Martínez, G. R., & Alvarez, E. D. (2021). Prevalence of diabetes mellitus in Spain in 2016 according to the Primary Care Clinical Database (BDCAP). Endocrinología, Diabetes y Nutrición (English Ed.), 68(2), 109–115. https://doi.org/10.1016/j.endien.2019.12.009

Menzie, D. E., & Dutta, S. (1989). Dispersivity as an oil reservoir rock characteristic (DOE/BC/10851-15, 5341564; p. DOE/BC/10851-15, 5341564). https://doi.org/10.2172/5341564

Microsoft. (2021). Neural Network Intelligence (2.1.0) [Computer software]. https://github.com/microsoft/nni

Microsoft Research. (2020). EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. (0.89).

Miller, O. L., Putman, A. L., Alder, J., Miller, M., Jones, D. K., & Wise, D. R. (2021). Changing climate drives future streamflow declines and challenges in meeting water demand across the southwestern United States. Journal of Hydrology X, 11, 100074. https://doi.org/10.1016/j.hydroa.2021.100074

Mittal, R., Ni, R., & Seo, J.-H. (2020). The flow physics of COVID-19. Journal of Fluid Mechanics, 894. https://doi.org/10.1017/jfm.2020.330

Motlhale, M., & Ncayiyana, J. R. (2019). Migration status and prevalence of diabetes and hypertension in Gauteng province, South Africa: Effect modification by demographic and socioeconomic characteristics—a cross-sectional population-based study. BMJ Open, 9(9), e027427. https://doi.org/10.1136/bmjopen-2018-027427

Moura, A. M., Martins, S. O., & Raposo, J. F. (2021). Consumption of antidiabetic medicines in Portugal: Results of a temporal data analysis of a thirteen-year study (2005–2017). BMC Endocrine Disorders, 21(1), 30. https://doi.org/10.1186/s12902-021-00686-w

NERVTAG. (2021). NERVTAG paper on COVID-19 variant of concern B.1.1.7 (p. 9). New and Emerging Respiratory Virus Threats Advisory Group. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/att

achment_data/file/961037/NERVTAG_note_on_B.1.1.7_severity_for_SAGE_77__1_.pdf

Ng, K., Alygizakis, N., Nika, M.-C., Galani, A., Oswald, P., Oswaldova, M., Čirka, Ľ., Kunkel, U., Macherius, A., Sengl, M., Mariani, G., Tavazzi, S., Skejo, H., Gawlik, B. M., Thomaidis, N. S., & Slobodnik, J. (2023). Wide-scope target screening characterization of legacy and emerging contaminants in the Danube River Basin by liquid and gas chromatography coupled with high-resolution mass spectrometry. Water Research, 230, 119539. https://doi.org/10.1016/j.watres.2022.119539

Nguyen, D., Hautekiet, P., Berete, F., Braekman, E., Charafeddine, R., Demarest, S., Drieskens, S., Gisle, L., Hermans, L., Tafforeau, J., & Van Der Heyden, J. (2020). The Belgian health examination survey: Objectives, design and methods. Archives of Public Health, 78(1), 50. https://doi.org/10.1186/s13690-020-00428-9

Nguyen, K. H. (2018). Analysis of emerging environmental contaminations using advanced instrumental tools: Application to human and environmental exposure. University of Birminham.

Niemuth, N. J., Jordan, R., Crago, J., Blanksma, C., Johnson, R., & Klaper, R. D. (2015). Metformin exposure at environmentally relevant concentrations causes potential endocrine disruption in adult male fish: Metformin causes potential endocrine disruption in male fish. Environmental Toxicology and Chemistry, 34(2), 291–296. https://doi.org/10.1002/etc.2793

Niemuth, N. J., & Klaper, R. D. (2015). Emerging wastewater contaminant metformin causes intersex and reduced fecundity in fish. Chemosphere, 135, 38–45. https://doi.org/10/f7qjzk

Niemuth, N. J., & Klaper, R. D. (2018). Low-dose metformin exposure causes changes in expression of endocrine disruption-associated genes. Aquatic Toxicology, 195, 33–40. https://doi.org/10/gc2zfk

Norman Network. (2016). NORMAN List of Emerging Substances. https://www.norman-network.net/?q=node/81

OECD. (2021). Health at a Glance 2021. https://www.oecd-ilibrary.org/content/publication/ae3016b9-en

Oertel, R., Baldauf, J., & Rossmann, J. (2018). Development and validation of a hydrophilic interaction liquid chromatography-tandem mass spectrometry method for the quantification of the antidiabetic drug metformin and six others

pharmaceuticals in wastewater. Journal of Chromatography A, 1556, 73–80. https://doi.org/10/gdnq2z

Ogunbanwo, O. M., Kay, P., Boxall, A. B., Wilkinson, J., Sinclair, C. J., Shabi, R. A., Fasasi, A. E., Lewis, G. A., Amoda, O. A., & Brown, L. E. (2022). High Concentrations of Pharmaceuticals in a Nigerian River Catchment. Environmental Toxicology and Chemistry, 41(3), 551–558. https://doi.org/10.1002/etc.4879

Ogurtsova, K., Da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J. E., & Makaroff, L. E. (2017). IDF Diabetes Atlas: Global estimates for the prevalence of diabetes for 2015 and 2040. Diabetes Research and Clinical Practice, 128, 40–50. https://doi.org/10.1016/j.diabres.2017.03.024

Oldenkamp, R., Hoeks, S., Čengić, M., Barbarossa, V., Burns, E. E., Boxall, A. B. A., & Ragas, A. M. J. (2018). A High-Resolution Spatial Model to Predict Exposure to Pharmaceuticals in European Surface Waters: ePiE. Environmental Science & Technology, 52(21), 12494–12503. https://doi.org/10.1021/acs.est.8b03862

O'Toole, Á., Hill, V., Pybus, O., Watts, A., & Bogoch, I. (2021, February 4). Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2. https://virological.org/t/tracking-the-international-spread-of-sars-cov-2-lineages-b-1-1-7-and-b-1-351-501y-v2/592

Pan, B. (2018). Application of XGBoost algorithm in hourly PM2.5 concentration prediction. IOP Conference Series: Earth and Environmental Science, 113, 012127. https://doi.org/10.1088/1755-1315/113/1/012127

Pan, S. J., & Yang, Q. (2010). A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Parrott, J. L., Pacepavicius, G., Shires, K., Clarence, S., Khan, H., Gardiner, M., Sullivan, C., & Alaee, M. (2021). Fathead minnow exposed to environmentally relevant concentrations of metformin for one life cycle show no adverse effects. FACETS, 6, 998–1023. https://doi.org/10.1139/facets-2020-0106

Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J., & Tapper, N. J. (2011). Quantifying the influence of local meteorology on air quality using generalized additive models. Atmospheric Environment, 45(6), 1328–1336. https://doi.org/10.1016/j.atmosenv.2010.11.051

Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge University Press.

Pearl, J. (2014). Interpretation and identification of causal mediation. Psychological Methods, 19(4), 459–481. https://doi.org/10.1037/a0036434

Pearl, J., & Mackenzie, D. (2018). The book of why: The new science of cause and effect. Basic Books.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559–572. https://doi.org/10.1080/14786440109462720

Podgorski, J., & Berg, M. (2020). Global threat of arsenic in groundwater. Science, 368(6493), 845–850. https://doi.org/10.1126/science.aba1510

Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., Rich, S., Wang, M., Buchan, I. E., & Bian, J. (2020). Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence, 2(7), 369–375. https://doi.org/10.1038/s42256-020-0197-y

Qiu, Z., Zhao, S., Feng, X., & He, Y. (2020). Transfer learning method for plastic pollution evaluation in soil using NIR sensor. Science of The Total Environment, 740, 140118. https://doi.org/10.1016/j.scitotenv.2020.140118

Qu, G., Li, X., Hu, L., & Jiang, G. (2020). An Imperative Need for Research on the Role of Environmental Factors in Transmission of Novel Coronavirus (COVID-19). Environmental Science & Technology, 54(7), 3730–3732. https://doi.org/10.1021/acs.est.0c01102

Rahman, Md. S., Azad, Md. A. K., Hasanuzzaman, Md., Salam, R., Islam, A. R. Md. T., Rahman, Md. M., & Hoque, M. Md. M. (2020). How air quality and COVID-19 transmission change under different lockdown scenarios? A case from Dhaka city, Bangladesh. Science of The Total Environment, 143161. https://doi.org/10/ghqh65

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

Rao, P. S. C., Davidson, J. M., Jessup, R. E., & Selim, H. M. (1979). Evaluation of Conceptual Models for Describing Nonequilibrium Adsorption-Desorption of Pesticides During Steady-flow in Soils. Soil Science Society of America

Journal, 43(1), 22–28. https://doi.org/10.2136/sssaj1979.03615995004300010004x

Rodríguez-Liébana, J. A., Mingorance, M. D., & Peña, A. (2018). Thiacloprid adsorption and leaching in soil: Effect of the composition of irrigation solutions. Science of The Total Environment, 610–611, 367–376. https://doi.org/10.1016/j.scitotenv.2017.08.028

Rolnick, D., Donti, P. L., Kaack, L. H., Kochanski, K., Lacoste, A., Sankaran, K., Ross, A. S., Milojevic-Dupont, N., Jaques, N., Waldman-Brown, A., Luccioni, A. S., Maharaj, T., Sherwin, E. D., Mukkavilli, S. K., Kording, K. P., Gomes, C. P., Ng, A. Y., Hassabis, D., Platt, J. C., … Bengio, Y. (2023). Tackling Climate Change with Machine Learning. ACM Computing Surveys, 55(2), 1–96. https://doi.org/10.1145/3485128

Romero, R., Erez, O., Hüttemann, M., Maymon, E., Panaitescu, B., Conde-Agudelo, A., Pacora, P., Yoon, B. H., & Grossman, L. I. (2017). Metformin, the aspirin of the 21st century: Its role in gestational diabetes mellitus, prevention of preeclampsia and cancer, and the promotion of longevity. American Journal of Obstetrics and Gynecology, 217(3), 282–302. https://doi.org/10.1016/j.ajog.2017.06.003

Rosario, D. K. A., Mutz, Y. S., Bernardes, P. C., & Conte-Junior, C. A. (2020). Relationship between COVID-19 and weather: Case study in a tropical country. International Journal of Hygiene and Environmental Health, 229, 113587. https://doi.org/10.1016/j.ijheh.2020.113587

Rostvall, A., Zhang, W., Dürig, W., Renman, G., Wiberg, K., Ahrens, L., & Gago-Ferrero, P. (2018). Removal of pharmaceuticals, perfluoroalkyl substances and other micropollutants from wastewater using lignite, Xylit, sand, granular activated carbon (GAC) and GAC+Polonite® in column tests – Role of physicochemical properties. Water Research, 137, 97–106. https://doi.org/10.1016/j.watres.2018.03.008

Rothan, H. A., & Byrareddy, S. N. (2020). The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. Journal of Autoimmunity, 109, 102433. https://doi.org/10.1016/j.jaut.2020.102433

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688–701. https://doi.org/10.1037/h0037350

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., & Nolan, G. P. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science, 308(5721), 523–529. https://doi.org/10.1126/science.1105809

Sadutto, D., Andreu, V., Ilo, T., Akkanen, J., & Picó, Y. (2021). Pharmaceuticals and personal care products in a Mediterranean coastal wetland: Impact of anthropogenic and spatial factors and environmental risk assessment. Environmental Pollution, 271, 116353. https://doi.org/10.1016/j.envpol.2020.116353

Sahadew, N., Pillay, S., & Singaram, V. (2022). Diabetes in the public healthcare sector of four South African provinces: A comparative analysis. South African Medical Journal, 855–859. https://doi.org/10.7196/SAMJ.2022.v112i11.16546

Sajedi-Hosseini, F., Malekian, A., Choubin, B., Rahmati, O., Cipullo, S., Coulon, F., & Pradhan, B. (2018). A novel machine learning-based approach for the risk assessment of nitrate groundwater contamination. Science of The Total Environment, 644, 954–962. https://doi.org/10.1016/j.scitotenv.2018.07.054

Sanborn, P., Lamontagne, L., & Hendershot, W. (2011). Podzolic soils of Canada: Genesis, distribution, and classification. Canadian Journal of Soil Science, 91(5), 843–880. https://doi.org/10.4141/cjss10024

Sarkodie, S. A., & Owusu, P. A. (2020). Impact of meteorological factors on COVID-19 pandemic: Evidence from top 20 countries with confirmed cases. Environmental Research, 191, 110101. https://doi.org/10.1016/j.envres.2020.110101

Sayyad, G., Price, G. W., Sharifi, M., & Khosravi, K. (2017). Fate and transport modeling of phthalate esters from biosolid amended soil under corn cultivation. Journal of Hazardous Materials, 323, 264–273. https://doi.org/10.1016/j.jhazmat.2016.07.032

Scheen, A. J. (2020). Metformin and COVID-19: From cellular mechanisms to reduced mortality. Diabetes & Metabolism, 46(6), 423–426. https://doi.org/10.1016/j.diabet.2020.07.006

Scheurer, M., Michel, A., Brauch, H.-J., Ruck, W., & Sacher, F. (2012). Occurrence and fate of the antidiabetic drug metformin and its metabolite guanylurea in the environment and during drinking water treatment. Water Research, 46(15), 4790–4802. https://doi.org/10/f37nr9

Schulze-Makuch, D. (2005). Longitudinal dispersivity data and implications for scaling behavior. Groundwater, 43(3), 443–456. https://doi.org/10.1111/j.1745-6584.2005.0051.x

Schwartz, H., Marushka, L., Chan, H. M., Batal, M., Sadik, T., Ing, A., Fediuk, K., & Tikhonov, C. (2021). Pharmaceuticals in source waters of 95 First Nations in Canada. Canadian Journal of Public Health, 112(1), 133–153. https://doi.org/10.17269/s41997-021-00499-3

Selim, H. M., Davidson, J. M., & Rao, P. S. C. (1977). Transport of Reactive Solutes through Multilayered Soils. Soil Science Society of America Journal, 41(1), 3–10. https://doi.org/10/dgbwg7

Shao, X.-T., Cong, Z.-X., Liu, S.-Y., Wang, Z., Zheng, X.-Y., & Wang, D.-G. (2021). Spatial analysis of metformin use compared with nicotine and caffeine consumption through wastewater-based epidemiology in China. Ecotoxicology and Environmental Safety, 208, 111623. https://doi.org/10.1016/j.ecoenv.2020.111623

Shao, X.-T., Zhao, Y.-T., Jiang, B., Li, Y.-Y., Lin, J.-G., & Wang, D.-G. (2023). Evaluation of Three Chronic Diseases by Selected Biomarkers in Wastewater. ACS ES&T Water, 3(4), 943–953. https://doi.org/10.1021/acsestwater.2c00452

Sharma, A., & Kiciman, E. (2020). DoWhy: An End-to-End Library for Causal Inference. arXiv:2011.04216 [Cs, Econ, Stat]. http://arxiv.org/abs/2011.04216

Shen, L., Zhao, T., Wang, H., Liu, J., Bai, Y., Kong, S., Zheng, H., Zhu, Y., & Shu, Z. (2021). Importance of meteorology in air pollution events during the city lockdown for COVID-19 in Hubei Province, Central China. Science of The Total Environment, 754, 142227. https://doi.org/10.1016/j.scitotenv.2020.142227

Shraim, A., Diab, A., Alsuhaimi, A., Niazy, E., Metwally, M., Amad, M., Sioud, S., & Dawoud, A. (2017). Analysis of some pharmaceuticals in municipal wastewater of Almadinah Almunawarah. Arabian Journal of Chemistry, 10, S719–S729. https://doi.org/10.1016/j.arabjc.2012.11.014

Simmons, C. T., Fenstemaker, T. R., & Sharp, J. M. (2001). Variable-density groundwater flow and solute transport in heterogeneous porous media: Approaches, resolutions and future challenges. Journal of Contaminant Hydrology, 52(1–4), 245–275. https://doi.org/10.1016/S0169-7722(01)00160-7

Šimůnek, J., & van Genuchten, M. Th. (2008). Modeling Nonequilibrium Flow and Transport Processes Using HYDRUS. Vadose Zone Journal, 7(2), 782–797. https://doi.org/10.2136/vzj2007.0074

Singha, S., Pasupuleti, S., Singha, S. S., Singh, R., & Kumar, S. (2021). Prediction of groundwater quality using efficient machine learning technique. Chemosphere, 276, 130265. https://doi.org/10.1016/j.chemosphere.2021.130265

Song, X., Lye, L. M., Chen, B., & Zhang, B. (2019). Differentiation of weathered chemically dispersed oil from weathered crude oil. Environmental Monitoring and Assessment, 191(5), 270. https://doi.org/10.1007/s10661-019-7392-5

Song, X.-B., Shao, X.-T., Liu, S.-Y., Tan, D.-Q., Wang, Z., & Wang, D.-G. (2020). Assessment of metformin, nicotine, caffeine, and methamphetamine use during Chinese public holidays. Chemosphere, 258, 127354. https://doi.org/10.1016/j.chemosphere.2020.127354

Soppi, A., Heino, P., Kurko, T., Maljanen, T., Saastamoinen, L., & Aaltonen, K. (2018). Growth of diabetes drug expenditure decomposed—A nationwide analysis. Health Policy, 122(12), 1326–1332. https://doi.org/10.1016/j.healthpol.2018.09.008

Srivastava, A. (2021). COVID-19 and air pollution and meteorology-an intricate relationship: A review. Chemosphere, 263, 128297. https://doi.org/10.1016/j.chemosphere.2020.128297

Steinhaus, H. (1956). Sur la division des corps mat´eriels en parties. Bull. Acad. Polon. Sci., IV (C1.III)(12), 801–804.

Sunyer, J., Dadvand, P., Foraster, M., Gilliland, F., & Nawrot, T. (2021). Environment and the COVID-19 pandemic. Environmental Research, 195, 110819. https://doi.org/10.1016/j.envres.2021.110819

Suryadi, Chew, L. Y., & Ong, Y.-S. (2023). Granger causality using Jacobian in neural networks. Chaos: An Interdisciplinary Journal of Nonlinear Science, 33(2), 023126. https://doi.org/10.1063/5.0106666

Tamayo, T., Brinks, R., Hoyer, A., Kuß, O., & Rathmann, W. (2016). The Prevalence and Incidence of Diabetes in Germany: An Analysis of Statutory Health Insurance Data on 65 Million Individuals From the Years 2009 and 2010. Deutsches Ärzteblatt International. https://doi.org/10.3238/arztebl.2016.0177

Tanabe, S., & Ramu, K. (2012). Monitoring temporal and spatial trends of legacy and emerging contaminants in marine environment: Results from the environmental specimen bank (es-BANK) of Ehime University, Japan. Marine Pollution Bulletin, 64(7), 1459–1474. https://doi.org/10.1016/j.marpolbul.2012.05.013

Tang, F. H. M., Lenzen, M., McBratney, A., & Maggi, F. (2021). Risk of pesticide pollution at the global scale. Nature Geoscience, 14(4), 206–210. https://doi.org/10.1038/s41561-021-00712-5

Tao, Y., Chen, B., Zhang, B. (Helen), Zhu, Z. (Joy), & Cai, Q. (2018). Occurrence, Impact, Analysis and Treatment of Metformin and Guanylurea in Coastal Aquatic Environments of Canada, USA and Europe. In Advances in Marine Biology (Vol. 81, pp. 23–58). Elsevier. https://doi.org/10.1016/bs.amb.2018.09.005

Tartakovsky, A. M., Marrero, C. O., Perdikaris, P., Tartakovsky, G. D., & Barajas-Solano, D. (2020). Physics-Informed Deep Neural Networks for Learning Parameters and Constitutive Relationships in Subsurface Flow Problems. Water Resources Research, 56(5). https://doi.org/10.1029/2019WR026731

Thompson, N. (2021, February 19). More contagious U.K. COVID-19 variant now found in all 10 provinces. Global News. https://globalnews.ca/news/7640125/coronavirus-canada-update-feb-13/

Thorndike, R. L. (1953). Who belongs in the family? Psychometrika, 18(4), 267–276. https://doi.org/10.1007/BF02289263

Tian, X., An, C., Chen, Z., & Tian, Z. (2021). Assessing the impact of COVID-19 pandemic on urban transportation and air quality in Canada. Science of The Total Environment, 765, 144270. https://doi.org/10/ghrkkd

Tisler, S., & Zwiener, C. (2018). Formation and occurrence of transformation products of metformin in wastewater and surface water. Science of The Total Environment, 628–629, 1121–1129. https://doi.org/10.1016/j.scitotenv.2018.02.105

Topor-Madry, R., Wojtyniak, B., Strojek, K., Rutkowski, D., Bogusławski, S., Ignaszewska-Wyrzykowska, A., Jarosz-Chobot, P., Czech, M., Kozierkiewicz, A., Chlebus, K., Jędrzejczyk, T., Mysliwiec, M., Polanska, J., Wysocki, M. J., & Zdrojewski, T. (2019). Prevalence of diabetes in Poland: A combined analysis of national databases. Diabetic Medicine, 36(10), 1209–1216. https://doi.org/10.1111/dme.13949

Trautwein, C., Berset, J.-D., Wolschke, H., & Kümmerer, K. (2014). Occurrence of the antidiabetic drug Metformin and its ultimate transformation product Guanylurea in several compartments of the aquatic cycle. Environment International, 70, 203–212. https://doi.org/10/f6ccr4

Uloko, A. E., Musa, B. M., Ramalan, M. A., Gezawa, I. D., Puepet, F. H., Uloko, A. T., Borodo, M. M., & Sada, K. B. (2018). Prevalence and Risk Factors for Diabetes Mellitus in Nigeria: A Systematic Review and Meta-Analysis. Diabetes Therapy, 9(3), 1307–1316. https://doi.org/10.1007/s13300-018-0441-1

Um, K., Hall, E. J., Katsoulakis, M. A., & Tartakovsky, D. M. (2019). Causality and Bayesian Network PDEs for multiscale representations of porous media. Journal of Computational Physics, 394, 658–678. https://doi.org/10.1016/j.jcp.2019.06.007

Ussery, E., Bridges, K. N., Pandelides, Z., Kirkwood, A. E., Guchardi, J., & Holdway, D. (2019). Developmental and Full-Life Cycle Exposures to Guanylurea and Guanylurea–Metformin Mixtures Results in Adverse Effects on Japanese Medaka (Oryzias latipes). Environmental Toxicology and Chemistry, 38(5), 1023–1028. https://doi.org/10.1002/etc.4403

van Doremalen, N., Bushmaker, T., Morris, D. H., Holbrook, M. G., Gamble, A., Williamson, B. N., Tamin, A., Harcourt, J. L., Thornburg, N. J., Gerber, S. I., Lloyd-Smith, J. O., de Wit, E., & Munster, V. J. (2020). Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. New England Journal of Medicine, 382(16), 1564–1567. https://doi.org/10.1056/NEJMc2004973

van Dyk, D. A., & Meng, X.-L. (2001). The Art of Data Augmentation. Journal of Computational and Graphical Statistics, 10(1), 1–50. https://doi.org/10.1198/10618600152418584

van Genuchten, M. Th. (1980). A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils. Soil Science Society of America Journal, 44(5), 892–898. https://doi.org/10.2136/sssaj1980.03615995004400050002x

van Genuchten, M. Th., & Wagenet, R. J. (1989). Two-Site/Two-Region Models for Pesticide Transport and Degradation: Theoretical Development and Analytical Solutions. Soil Science Society of America Journal, 53(5), 1303–1310. https://doi.org/10.2136/sssaj1989.03615995005300050001x

van Nuijs, A. L. N., Tarcomnicu, I., Simons, W., Bervoets, L., Blust, R., Jorens, P. G., Neels, H., & Covaci, A. (2010). Optimization and validation of a

hydrophilic interaction liquid chromatography–tandem mass spectrometry method for the determination of 13 top-prescribed pharmaceuticals in influent wastewater. Analytical and Bioanalytical Chemistry, 398(5), 2211–2222. https://doi.org/10.1007/s00216-010-4101-1

Varian, H. R. (2016). Causal inference in economics and marketing. Proceedings of the National Academy of Sciences, 113(27), 7310–7315. https://doi.org/10.1073/pnas.1510479113

Vörösmarty, C. J., Green, P., Salisbury, J., & Lammers, R. B. (2000). Global Water Resources: Vulnerability from Climate Change and Population Growth. Science, 289(5477), 284–288. https://doi.org/10.1126/science.289.5477.284

Walensky, R. P., Walke, H. T., & Fauci, A. S. (2021). SARS-CoV-2 Variants of Concern in the United States—Challenges and Opportunities. JAMA. https://doi.org/10.1001/jama.2021.2294

Wang, J., Xu, X., Wang, S., He, S., & He, P. (2021). Heterogeneous effects of COVID-19 lockdown measures on air quality in Northern China. Applied Energy, 282, 116179. https://doi.org/10.1016/j.apenergy.2020.116179

Wang, M., Liu, F., & Zheng, M. (2020). Air quality improvement from COVID-19 lockdown: Evidence from China. Air Quality, Atmosphere & Health. https://doi.org/10.1007/s11869-020-00963-y

Wang, S., Wasswa, J., Feldman, A. C., Kabenge, I., Kiggundu, N., & Zeng, T. (2022). Suspect screening to support source identification and risk assessment of organic micropollutants in the aquatic environment of a Sub-Saharan African urban center. Water Research, 220, 118706. https://doi.org/10.1016/j.watres.2022.118706

Wang, Y., Shi, L., Hu, X., Song, W., & Wang, L. (2023). Multiphysics-Informed Neural Networks for Coupled Soil Hydrothermal Modeling. Water Resources Research, 59(1), e2022WR031960. https://doi.org/10.1029/2022WR031960

Wang, Y., Wen, Y., Wang, Y., Zhang, S., Zhang, K. M., Zheng, H., Xing, J., Wu, Y., & Hao, J. (2020). Four-Month Changes in Air Quality during and after the COVID-19 Lockdown in Six Megacities in China. Environmental Science & Technology Letters, 7(11), 802–808. https://doi.org/10/ghqf24

Wei, Q., Zhou, K., Chen, J., Zhang, Q., Lu, T., Farooq, U., Chen, W., Li, D., & Qi, Z. (2021). Insights into the molecular mechanism of tetracycline transport in saturated porous media affected by low-molecular-weight organic acids: Role

of the functional groups and molecular size. Science of The Total Environment, 799, 149361. https://doi.org/10.1016/j.scitotenv.2021.149361

WHO. (2024). Anatomical Therapeutic Chemical (ATC) classification system and the Defined Daily Dose (DDD). https://www.who.int/teams/health-product-and-policy-standards/inn/atc-ddd

Wilkinson, J. L., Boxall, A. B. A., Kolpin, D. W., Leung, K. M. Y., Lai, R. W. S., Galbán-Malagón, C., Adell, A. D., Mondon, J., Metian, M., Marchant, R. A., Bouzas-Monroy, A., Cuni-Sanchez, A., Coors, A., Carriquiriborde, P., Rojo, M., Gordon, C., Cara, M., Moermond, M., Luarte, T., … Teta, C. (2022). Pharmaceutical pollution of the world's rivers. Proceedings of the National Academy of Sciences, 119(8), e2113947119. https://doi.org/10.1073/pnas.2113947119

World Health Organization. (2020). Advice on the use of masks in the context of COVID-19: Interim guidance (Geneva). Geneva: World Health Organization. https://apps.who.int/iris/handle/10665/332293

Xiao, Y., Shao, X.-T., Tan, D.-Q., Yan, J.-H., Pei, W., Wang, Z., Yang, M., & Wang, D.-G. (2019). Assessing the trend of diabetes mellitus by analyzing metformin as a biomarker in wastewater. Science of The Total Environment, 688, 281–287. https://doi.org/10.1016/j.scitotenv.2019.06.117

Xin, X., Huang, G., An, C., & Feng, R. (2019). Interactive Toxicity of Triclosan and Nano-TiO2 to Green Alga Eremosphaera viridis in Lake Erie: A New Perspective Based on Fourier Transform Infrared Spectromicroscopy and Synchrotron-Based X-ray Fluorescence Imaging. Environmental Science & Technology, 53(16), 9884–9894. https://doi.org/10/ggf2s5

Xin, X., Huang, G., An, C., Raina-Fulton, R., & Weger, H. (2019). Insights into Long-Term Toxicity of Triclosan to Freshwater Green Algae in Lake Erie. Environmental Science & Technology, 53(4), 2189–2198. https://doi.org/10/ggf2s6

Xing, Y., Chen, X., Wagner, R. E., Zhuang, J., & Chen, X. (2020). Coupled effect of colloids and surface chemical heterogeneity on the transport of antibiotics in porous media. Science of The Total Environment, 713, 136644. https://doi.org/10.1016/j.scitotenv.2020.136644

Xiong, R., Zheng, Y., Chen, N., Tian, Q., Liu, W., Han, F., Jiang, S., Lu, M., & Zheng, Y. (2022). Predicting Dynamic Riverine Nitrogen Export in Unmonitored Watersheds: Leveraging Insights of AI from Data-Rich Regions.

Environmental Science & Technology, 56(14), 10530–10542. https://doi.org/10.1021/acs.est.2c02232

Xu, B., Wang, N., Chen, T., & Li, M. (2015). Empirical Evaluation of Rectified Activations in Convolutional Network (arXiv:1505.00853). arXiv. http://arxiv.org/abs/1505.00853

Xu, K., Cui, K., Young, L.-H., Hsieh, Y.-K., Wang, Y.-F., Zhang, J., & Wan, S. (2020). Impact of the COVID-19 Event on Air Quality in Central China. Aerosol and Air Quality Research, 20(5), 915–929. https://doi.org/10.4209/aaqr.2020.04.0150

Xue, L. K., Wang, T., Gao, J., Ding, A. J., Zhou, X. H., Blake, D. R., Wang, X. F., Saunders, S. M., Fan, S. J., Zuo, H. C., Zhang, Q. Z., & Wang, W. X. (2014). Ground-level ozone in four Chinese cities: Precursors, regional transport and heterogeneous processes. Atmospheric Chemistry and Physics, 14(23), 13175–13188. https://doi.org/10.5194/acp-14-13175-2014

Yan, J.-H., Xiao, Y., Tan, D.-Q., Shao, X.-T., Wang, Z., & Wang, D.-G. (2019). Wastewater analysis reveals spatial pattern in consumption of anti-diabetes drug metformin in China. Chemosphere, 222, 688–695. https://doi.org/10.1016/j.chemosphere.2019.01.151

Yang, D., Zheng, Q., Thai, P., Ahmed, F., O'Brien, J. W., Mueller, J. F., Thomas, K. V., & Tscharke, B. (2022). A nationwide wastewater-based assessment of metformin consumption across Australia. Environment International, 107282. https://doi.org/10.1016/j.envint.2022.107282

Yao, B., Yan, S., Lian, L., Yang, X., Wan, C., Dong, H., & Song, W. (2018). Occurrence and indicators of pharmaceuticals in Chinese streams: A nationwide study. Environmental Pollution, 236, 889–898. https://doi.org/10/gdc9qj

Ye, X., Chen, B., Jing, L., Zhang, B., & Liu, Y. (2019). Multi-agent hybrid particle swarm optimization (MAHPSO) for wastewater treatment network planning. Journal of Environmental Management, 234, 525–536. https://doi.org/10/gf2kmv

Ye, X., Chen, B., Lee, K., Storesund, R., & Zhang, B. (2020). An integrated offshore oil spill response decision making approach by human factor analysis and fuzzy preference evaluation. Environmental Pollution, 262, 114294. https://doi.org/10.1016/j.envpol.2020.114294

You, K., & Zhan, H. (2013). New solutions for solute transport in a finite column with distance-dependent dispersivities and time-dependent solute sources.

Journal of Hydrology, 487, 87–97. https://doi.org/10.1016/j.jhydrol.2013.02.027

Younes, A., Fahs, M., Ataie-Ashtiani, B., & Simmons, C. T. (2020). Effect of distance-dependent dispersivity on density-driven flow in porous media. Journal of Hydrology, 589, 125204. https://doi.org/10.1016/j.jhydrol.2020.125204

Zakari, S., Liu, H., Tong, L., Wang, Y., & Liu, J. (2016). Transport of bisphenol-A in sandy aquifer sediment: Column experiment. Chemosphere, 144, 1807–1814. https://doi.org/10.1016/j.chemosphere.2015.10.081

Zakari, S., Liu, H., & Zhou, H. (2019). Transport velocities of aniline and nitrobenzene in sandy sediment. Journal of Soils and Sediments, 19(5), 2570–2579. https://doi.org/10.1007/s11368-019-02287-6

Zamani Joharestani, M., Cao, C., Ni, X., Bashir, B., & Talebiesfandarani, S. (2019). PM2.5 Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data. Atmosphere, 10(7), Article 7. https://doi.org/10.3390/atmos10070373

Zhang, R., He, Y., Yao, L., Chen, J., Zhu, S., Rao, X., Tang, P., You, J., Hua, G., Zhang, L., Ju, F., & Wu, L. (2021). Metformin chlorination byproducts in drinking water exhibit marked toxicities of a potential health concern. Environment International, 146, 106244. https://doi.org/10.1016/j.envint.2020.106244

Zhang, X., Tang, M., Guo, F., Wei, F., Yu, Z., Gao, K., Jin, M., Wang, J., & Chen, K. (2021). Associations between air pollution and COVID-19 epidemic during quarantine period in China. Environmental Pollution, 268, 115897. https://doi.org/10.1016/j.envpol.2020.115897

Zheng, Q., Du, P., Wang, Z., Zhang, L., Zhu, Z., Huang, J., Wang, Z., Hall, W., Dang, A. K., Wang, D., Li, X., & Thai, P. K. (2023). Nation-Wide Wastewater-Based Epidemiology Assessment of Metformin Usage in China: 2014–2020. ACS ES&T Water, 3(1), 195–202. https://doi.org/10.1021/acsestwater.2c00489

Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., Ma, X., Marrone, B. L., Ren, Z. J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B. M., Xiao, X., Yu, X., Zhu, J.-J., & Zhang, H. (2021). Machine Learning: New Ideas and Tools in Environmental Science and Engineering. Environmental Science & Technology, acs.est.1c01339. https://doi.org/10.1021/acs.est.1c01339

Zhou, L., Martin, S., Cheng, W., Lassabatere, L., Boily, J.-F., & Hanna, K. (2019). Water Flow Variability Affects Adsorption and Oxidation of Ciprofloxacin onto Hematite. Environmental Science & Technology, 53(17), 10102–10109. https://doi.org/10.1021/acs.est.9b03214

Zhu, J., Yang, M., & Ren, Z. J. (2023). Machine Learning in Environmental Research: Common Pitfalls and Best Practices. Environmental Science & Technology, 57(46), 17671–17689. https://doi.org/10.1021/acs.est.3c00026

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine, 382(8), 727–733. https://doi.org/10.1056/NEJMoa2001017

Zou, X., Guo, H., Jiang, C., Nguyen, D. V., Chen, G.-H., & Wu, D. (2023). Physics-informed neural network-based serial hybrid model capturing the hidden kinetics for sulfur-driven autotrophic denitrification process. Water Research, 243, 120331. https://doi.org/10.1016/j.watres.2023.120331

APPENDICES

Appendix A EnvCausal Framework Benchmark

SCM were further applied on two public datasets: Infant Health and Development Program (IHDP; Louizos et al., 2017) Dataset and Lanlode Dataset (Dehejia & Wahba, 1999). The IHDP dataset (n=747) is from a randomized experiment that began in 1985 targeting infant health, which means that the ground truth of the causal relationship in the dataset is known. The dataset consists of measurements on the child (birth weight, head circumference, weeks born preterm, birth order, first born, neonatal health, index sex, twin status) as well as mother status and behaviours during the pregnancy (consumption status of cigarettes, alcohol and drugs, age, marital status, educational attainment, employment, prenatal care, family residing site). The treatment variable in the dataset is if the infant received both intensive high-quality child care and home visits from a trained provider. LaLonde Dataset (n=445) is another well-known dataset that aims to investigate the effect of an employment training program, National Supported Work Demonstration (NSW), on wage increases (i.e., real income in 1978). Since the applicants were admitted randomly to the program, the ground truth within the dataset is also known as in the IHDP Dataset. The dataset also has the features such as age, years of schooling, indicator variables for race, martial status, high school diploma, real earnings in 1974 and 1975, and whether earnings in 1974 or 1975 being zero. The SCM identified causal relationships in both dataset, which passed three refutation methods (i.e., add random common cause, replace treatment with placebo, remove random subset of data). The estimates causal effects are 3.41 and 1614.16, respectively. A Jupyter notebook with the causal estimation and refutation results can be found in the GitHub repository of the study (https://github.com/kangqiao-ctrl/EnvCausal/tree/main/benchmark/hdpi_lalonde).

Structural Agnostic Model (SAM) was applied on another well-known dataset, Sachs Dataset (Sachs et al., 2005), which consists of simultaneous measurements of multiple phosphorylated proteins and phospholipid components in thousands of individual primary human immune system cells. The dataset was generated with molecular interventions which perturbed the cells. SAM was applied to the dataset to test capability in recovering the causal network. Two important metrics, Precision and Recall were calculated based on Equation S6 and S7:

$$Precision = \frac{TP}{TP+FP} \tag{A.1}$$

$$Recall = \frac{TP}{TP+FN} \tag{A.2}$$

Where TP is the number of true positives, FP is the number of false positives, FN is the number of false negatives.

For the Sachs dataset, SAM acquired an Area Under Precision-Recall Curve (AUPR) of 0.311. Since the corresponding baseline is 0.168 for this case, AUPR of such value is considered decent. A Jupyter notebook of the SAM benchmark on Sachs Dataset can be found in the GitHub repository of the study (https://github.com/kangqiao-ctrl/EnvCausal/tree/main/benchmark/sachs).

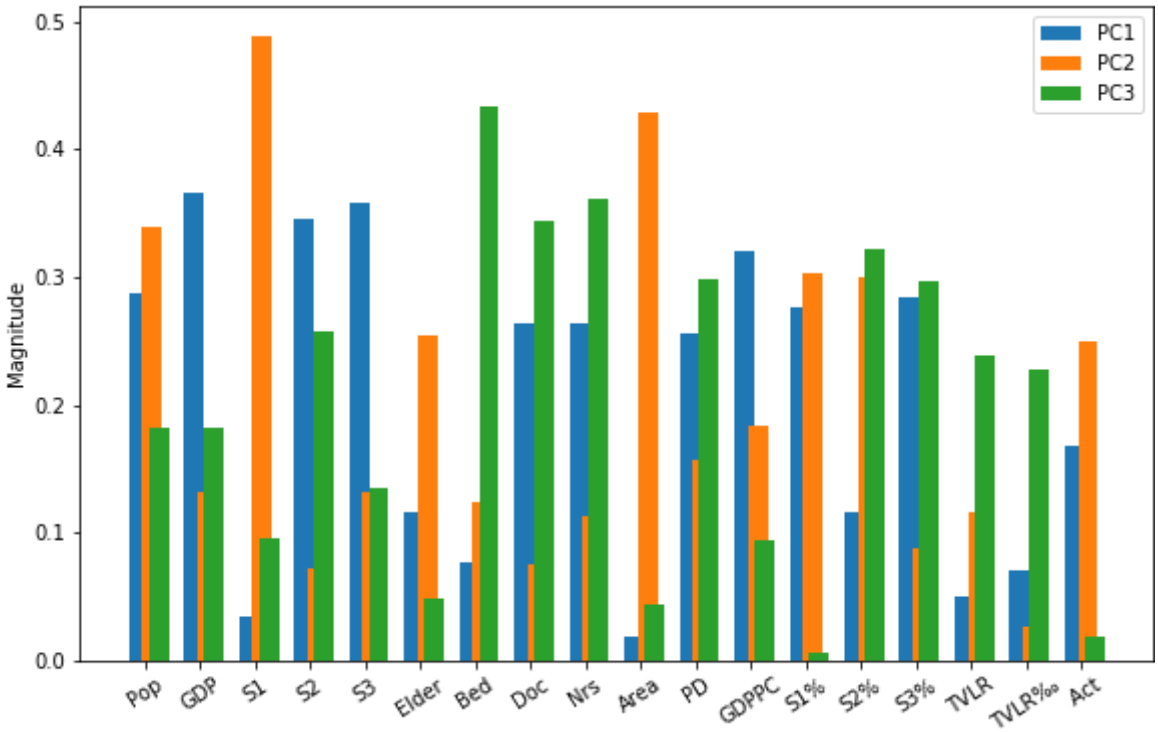


Figure A.1 Contribution of each feature to different principal components. Pop: population; S1/S2/S3: primary, secondary, tertiary sector of GDP; Elder: elderly population percentage (over 60-year-old); Bed/Doc/Nrs: hospital beds/ registered medical doctors/ registered nurses per thousand people; TVLR: travellers from Wuhan; TVLR‰: Wuhan travellers per thousand; Act: the average degree of activeness before the 2020 Spring Festival. Explained variance by PC1-3: 31.1%, 18.2%, 14.0%.

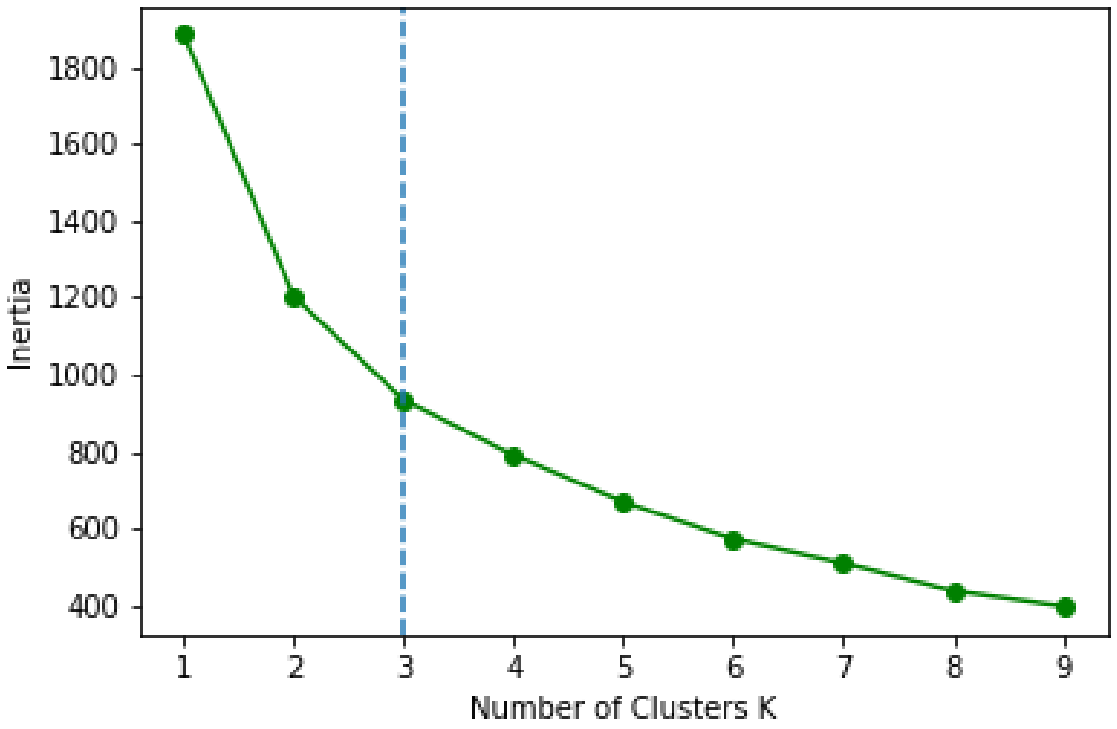


Figure A.2 Explained variance and number of clusters. The "elbow" is indicated by the blue dashed line. The number of clusters chosen should therefore be 3.

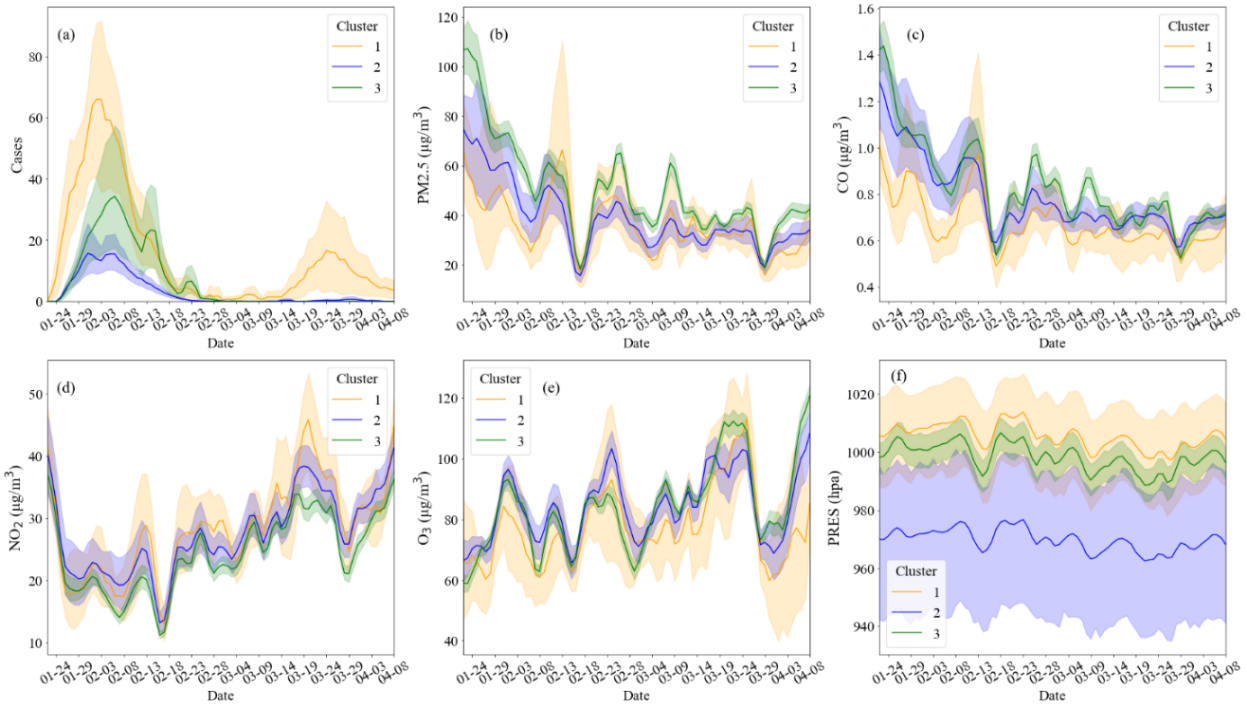
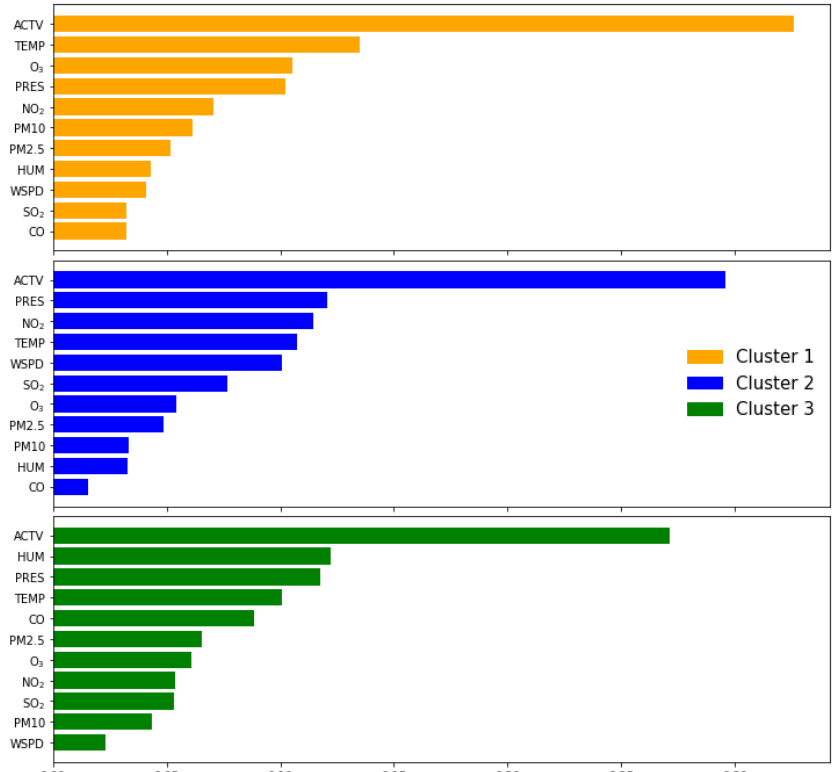


Figure A.3 COVID-19 cases (a) and five selected features: (b) PM2.5, (c) CO, (d) NO_2, (e) O_3, (f) atmospheric pressure. Colored bands indicate 95% confidence intervals.

(a)



(b)

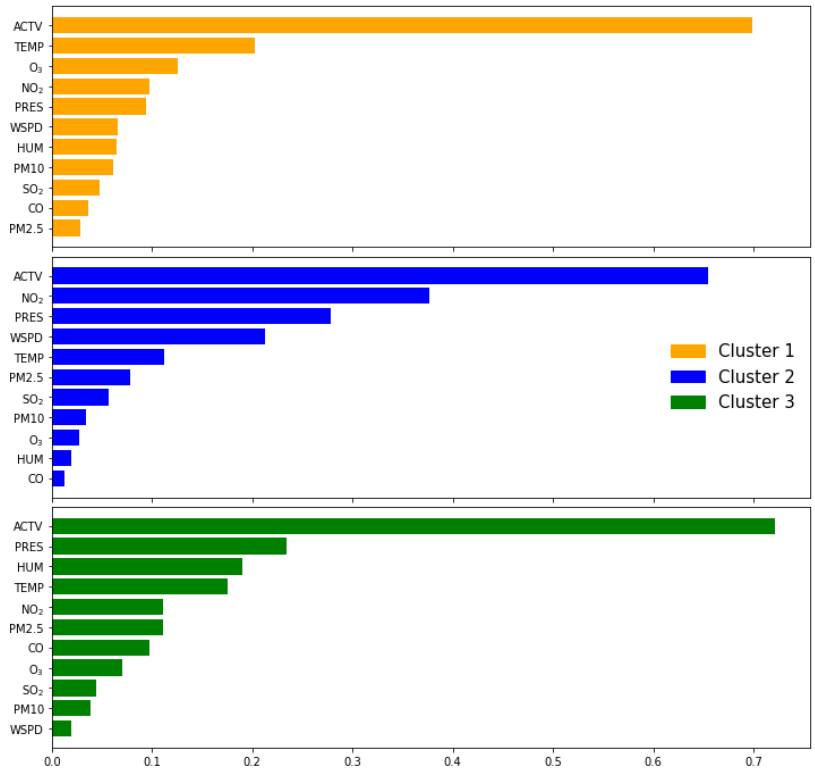


Figure A.4 Feature importance and ranking in different clusters with no "elapsed days" feature (*a) Normalized Total Gain, (b)Permutation Importance*

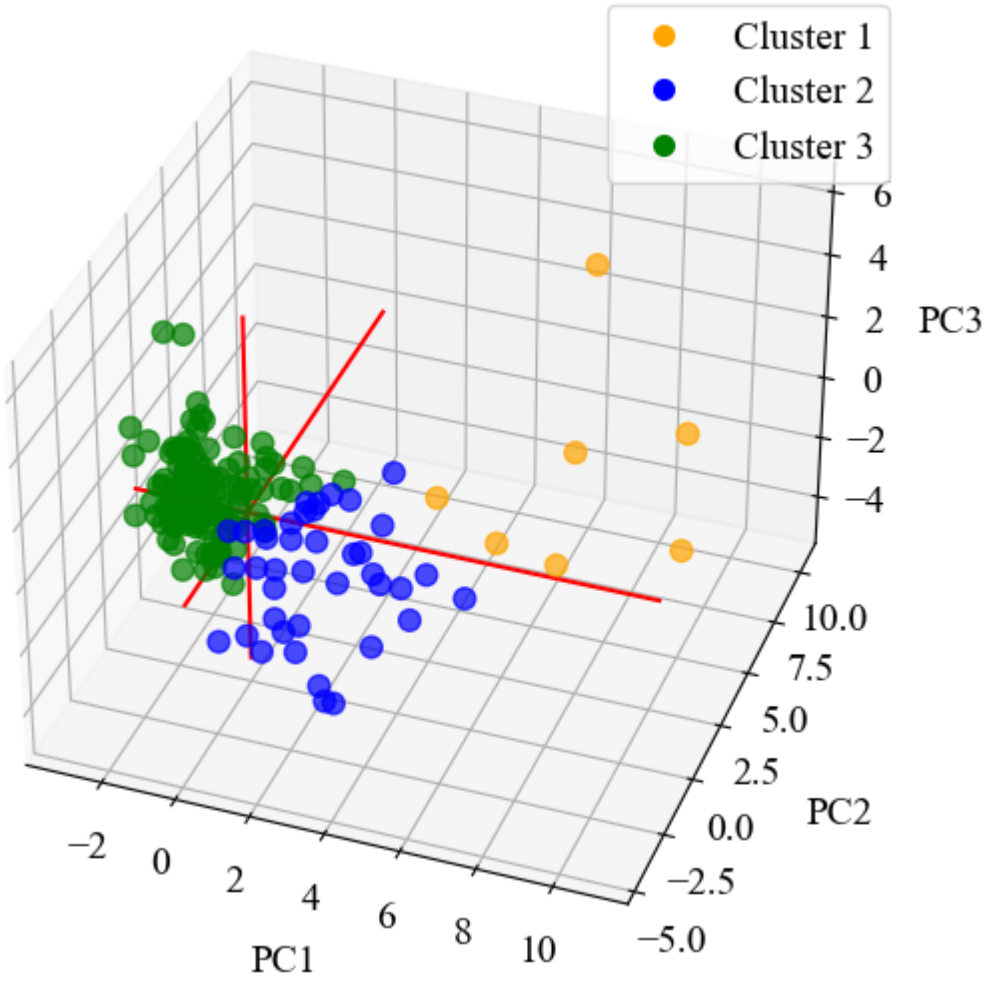


Figure A.5 Clustered cities in the principal component space.

Table A.1 Weighted adjacency matrix generated by SAM

Cluster 1 - Overall													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.818	0.004	0.596	0.240	0.035	0.010	0.005	0.250	0.539	0.230	0.312	0.415
PM10	0.260	0.000	0.411	0.008	0.547	0.452	0.335	0.358	0.255	0.026	0.150	0.031	0.011
SO_2	0.082	0.247	0.000	0.295	0.368	0.440	0.488	0.296	0.282	0.118	0.001	0.366	0.010
CO	0.634	0.012	0.547	0.000	0.730	0.316	0.677	0.007	0.466	0.020	0.323	0.300	0.543
NO_2	0.009	0.052	0.600	0.259	0.000	0.053	0.001	0.009	0.684	0.053	0.236	0.165	0.037
O_3	0.135	0.010	0.008	0.377	0.025	0.000	0.393	0.244	0.233	0.519	0.445	0.307	0.143
HMD	0.161	0.340	0.535	0.345	0.003	0.318	0.000	0.184	0.409	0.392	0.432	0.287	0.260
PRES	0.151	0.312	0.513	0.326	0.686	0.786	0.174	0.000	0.319	0.544	0.616	0.177	0.002
WSPD	0.004	0.002	0.313	0.178	0.383	0.001	0.267	0.010	0.000	0.001	0.012	0.517	0.001
TEMP	0.465	0.475	0.551	0.258	0.278	0.338	0.251	0.425	0.487	0.000	0.597	0.265	0.455
ACTV	0.014	0.016	0.046	0.015	0.455	0.024	0.002	0.003	0.600	0.083	0.000	0.330	0.515
CASES	0.031	0.144	0.003	0.143	0.268	0.002	0.234	0.167	0.326	0.073	0.014	0.000	0.127
DAYS	0.196	0.502	0.125	0.371	0.855	0.132	0.014	0.004	0.252	0.348	0.975	0.888	0.000
Cluster 1 – Spreading phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.696	0.043	0.553	0.036	0.057	0.004	0.006	0.158	0.110	0.045	0.045	0.043

PM10	0.051	0.000	0.179	0.045	0.037	0.500	0.021	0.028	0.095	0.074	0.005	0.002	0.012
SO_2	0.001	0.200	0.000	0.147	0.008	0.010	0.709	0.237	0.129	0.169	0.305	0.471	0.002
CO	0.117	0.001	0.415	0.000	0.372	0.002	0.007	0.001	0.092	0.015	0.004	0.006	0.010
NO_2	0.075	0.008	0.004	0.221	0.000	0.005	0.005	0.023	0.572	0.002	0.445	0.004	0.003
O_3	0.014	0.039	0.001	0.029	0.076	0.000	0.698	0.158	0.101	0.008	0.269	0.001	0.002
HMD	0.001	0.017	0.048	0.003	0.035	0.226	0.000	0.004	0.232	0.030	0.005	0.009	0.011
PRES	0.224	0.171	0.557	0.010	0.578	0.584	0.021	0.000	0.769	0.404	0.021	0.005	0.001
WSPD	0.002	0.013	0.218	0.008	0.235	0.005	0.074	0.002	0.000	0.012	0.003	0.031	0.004
TEMP	0.159	0.002	0.617	0.010	0.002	0.473	0.382	0.471	0.137	0.000	0.608	0.507	0.031
ACTV	0.017	0.146	0.027	0.043	0.235	0.409	0.628	0.010	0.130	0.003	0.000	0.014	0.031
CASES	0.004	0.006	0.019	0.053	0.093	0.003	0.691	0.065	0.381	0.095	0.062	0.000	0.174
DAYS	0.002	0.042	0.019	0.280	0.852	0.476	0.631	0.011	0.079	0.588	0.747	0.700	0.000
Cluster 1 – Post-peak phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.443	0.006	0.436	0.028	0.005	0.023	0.010	0.009	0.128	0.291	0.113	0.009
PM10	0.476	0.000	0.901	0.005	0.025	0.152	0.300	0.838	0.090	0.028	0.003	0.012	0.615
SO_2	0.138	0.016	0.000	0.056	0.034	0.347	0.267	0.070	0.632	0.016	0.171	0.009	0.052
CO	0.507	0.008	0.076	0.000	0.509	0.097	0.735	0.042	0.486	0.062	0.002	0.267	0.006
NO_2	0.005	0.004	0.503	0.217	0.000	0.059	0.004	0.508	0.881	0.031	0.477	0.394	0.051
O_3	0.003	0.006	0.167	0.469	0.300	0.000	0.517	0.112	0.001	0.214	0.029	0.571	0.045

HMD	0.037	0.127	0.250	0.080	0.009	0.138	0.000	0.238	0.629	0.022	0.001	0.318	0.012
PRES	0.245	0.020	0.525	0.148	0.280	0.426	0.222	0.000	0.444	0.379	0.896	0.730	0.002
WSPD	0.024	0.001	0.180	0.032	0.014	0.001	0.223	0.015	0.000	0.005	0.005	0.059	0.001
TEMP	0.585	0.125	0.561	0.621	0.075	0.446	0.741	0.270	0.314	0.000	0.579	0.033	0.071
ACTV	0.001	0.005	0.056	0.011	0.047	0.032	0.011	0.004	0.054	0.017	0.000	0.139	0.147
CASES	0.027	0.006	0.013	0.010	0.004	0.031	0.006	0.005	0.009	0.010	0.129	0.000	0.037
DAYS	0.008	0.176	0.171	0.507	0.517	0.148	0.438	0.029	0.895	0.379	0.681	0.735	0.000
Cluster 2 - Overall													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.718	0.199	0.785	0.263	0.026	0.477	0.402	0.471	0.192	0.097	0.155	0.178
PM10	0.445	0.000	0.152	0.402	0.320	0.334	0.459	0.207	0.008	0.005	0.003	0.008	0.332
SO_2	0.280	0.618	0.000	0.477	0.564	0.007	0.428	0.431	0.323	0.479	0.258	0.034	0.263
CO	0.310	0.277	0.426	0.000	0.355	0.393	0.224	0.303	0.084	0.357	0.001	0.273	0.455
NO_2	0.087	0.263	0.319	0.420	0.000	0.510	0.049	0.152	0.249	0.443	0.379	0.205	0.404
O_3	0.003	0.151	0.310	0.152	0.163	0.000	0.298	0.183	0.066	0.310	0.279	0.149	0.162
HMD	0.239	0.265	0.307	0.319	0.288	0.486	0.000	0.448	0.288	0.352	0.003	0.005	0.480
PRES	0.319	0.132	0.188	0.299	0.398	0.268	0.352	0.000	0.287	0.208	0.501	0.415	0.539
WSPD	0.001	0.167	0.270	0.538	0.162	0.200	0.294	0.446	0.000	0.017	0.256	0.466	0.316
TEMP	0.276	0.167	0.207	0.415	0.075	0.397	0.409	0.306	0.167	0.000	0.370	0.340	0.436
ACTV	0.293	0.397	0.179	0.026	0.399	0.416	0.016	0.092	0.241	0.309	0.000	0.546	0.405

CASES	0.308	0.450	0.001	0.352	0.029	0.144	0.201	0.390	0.152	0.087	0.127	0.000	0.596
DAYS	0.361	0.472	0.071	0.457	0.481	0.369	0.006	0.007	0.071	0.271	0.736	0.166	0.000
Cluster 2 – Spreading phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.812	0.008	0.514	0.354	0.017	0.573	0.279	0.254	0.523	0.005	0.068	0.221
PM10	0.082	0.000	0.250	0.409	0.409	0.010	0.051	0.018	0.032	0.224	0.022	0.088	0.004
SO_2	0.149	0.324	0.000	0.233	0.077	0.126	0.116	0.261	0.340	0.290	0.297	0.014	0.006
CO	0.133	0.007	0.457	0.000	0.359	0.381	0.048	0.040	0.314	0.203	0.030	0.018	0.403
NO_2	0.011	0.088	0.132	0.481	0.000	0.109	0.005	0.082	0.235	0.064	0.278	0.351	0.307
O_3	0.002	0.001	0.113	0.196	0.008	0.000	0.183	0.126	0.403	0.247	0.012	0.007	0.079
HMD	0.078	0.137	0.374	0.005	0.001	0.692	0.000	0.450	0.322	0.118	0.089	0.003	0.464
PRES	0.003	0.313	0.361	0.244	0.193	0.117	0.127	0.000	0.085	0.165	0.009	0.238	0.024
WSPD	0.068	0.002	0.124	0.112	0.005	0.186	0.002	0.184	0.000	0.003	0.058	0.294	0.012
TEMP	0.003	0.011	0.442	0.307	0.177	0.243	0.308	0.168	0.486	0.000	0.259	0.116	0.260
ACTV	0.009	0.121	0.254	0.002	0.176	0.131	0.002	0.148	0.256	0.336	0.000	0.331	0.386
CASES	0.207	0.161	0.004	0.008	0.029	0.190	0.421	0.405	0.423	0.087	0.004	0.000	0.545
DAYS	0.003	0.004	0.001	0.089	0.056	0.450	0.086	0.020	0.010	0.274	0.483	0.247	0.000
Cluster 2 – Post-peak phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.514	0.194	0.411	0.408	0.003	0.340	0.208	0.062	0.260	0.195	0.001	0.002

PM10	0.360	0.000	0.112	0.022	0.236	0.166	0.187	0.071	0.063	0.025	0.034	0.100	0.231
SO_2	0.069	0.456	0.000	0.158	0.239	0.002	0.272	0.173	0.173	0.431	0.225	0.351	0.227
CO	0.279	0.306	0.333	0.000	0.437	0.147	0.003	0.315	0.332	0.248	0.190	0.214	0.254
NO_2	0.057	0.313	0.384	0.250	0.000	0.007	0.062	0.250	0.257	0.107	0.212	0.235	0.195
O_3	0.135	0.079	0.046	0.228	0.065	0.000	0.189	0.279	0.171	0.245	0.010	0.047	0.247
HMD	0.232	0.271	0.254	0.303	0.212	0.426	0.000	0.411	0.259	0.306	0.078	0.147	0.003
PRES	0.276	0.105	0.273	0.244	0.185	0.196	0.278	0.000	0.317	0.169	0.467	0.422	0.092
WSPD	0.002	0.108	0.271	0.111	0.001	0.002	0.020	0.303	0.000	0.037	0.284	0.585	0.035
TEMP	0.128	0.235	0.229	0.363	0.005	0.555	0.299	0.447	0.126	0.000	0.088	0.159	0.380
ACTV	0.264	0.198	0.368	0.120	0.190	0.148	0.105	0.096	0.167	0.238	0.000	0.120	0.241
CASES	0.201	0.362	0.004	0.004	0.168	0.022	0.054	0.127	0.024	0.070	0.094	0.000	0.467
DAYS	0.002	0.070	0.008	0.107	0.376	0.220	0.115	0.023	0.023	0.250	0.435	0.214	0.000

Cluster 3 - Overall

	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.539	0.017	0.426	0.193	0.199	0.192	0.204	0.228	0.211	0.002	0.226	0.136
PM10	0.423	0.000	0.195	0.029	0.138	0.194	0.268	0.060	0.204	0.222	0.150	0.065	0.005
SO_2	0.229	0.259	0.000	0.231	0.249	0.092	0.312	0.243	0.190	0.208	0.209	0.006	0.024
CO	0.145	0.253	0.129	0.000	0.221	0.210	0.249	0.308	0.089	0.181	0.207	0.119	0.207
NO_2	0.091	0.212	0.215	0.212	0.000	0.209	0.100	0.126	0.207	0.272	0.144	0.069	0.179
O_3	0.206	0.238	0.295	0.140	0.171	0.000	0.206	0.189	0.227	0.227	0.277	0.121	0.131

HMD	0.202	0.205	0.238	0.135	0.243	0.361	0.000	0.190	0.238	0.210	0.207	0.148	0.210
PRES	0.005	0.290	0.106	0.202	0.150	0.227	0.185	0.000	0.181	0.252	0.111	0.165	0.092
WSPD	0.129	0.186	0.161	0.269	0.132	0.314	0.163	0.138	0.000	0.204	0.221	0.123	0.218
TEMP	0.206	0.228	0.223	0.166	0.160	0.221	0.176	0.211	0.182	0.000	0.166	0.071	0.200
ACTV	0.063	0.140	0.143	0.178	0.176	0.239	0.140	0.119	0.186	0.165	0.000	0.106	0.244
CASES	0.011	0.000	0.008	0.217	0.364	0.038	0.173	0.329	0.053	0.217	0.634	0.000	0.278
DAYS	0.197	0.204	0.317	0.179	0.330	0.226	0.012	0.111	0.131	0.193	0.157	0.099	0.000
Cluster 3 – Spreading phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.389	0.050	0.686	0.341	0.479	0.662	0.278	0.021	0.223	0.173	0.090	0.143
PM10	0.833	0.000	0.475	0.755	0.517	0.080	0.405	0.002	0.082	0.221	0.251	0.034	0.154
SO_2	0.277	0.380	0.000	0.334	0.734	0.057	0.452	0.264	0.130	0.291	0.209	0.001	0.010
CO	0.037	0.243	0.207	0.000	0.208	0.058	0.592	0.041	0.270	0.162	0.001	0.161	0.225
NO_2	0.058	0.214	0.436	0.391	0.000	0.288	0.253	0.002	0.034	0.008	0.048	0.016	0.022
O_3	0.132	0.004	0.145	0.334	0.143	0.000	0.282	0.158	0.284	0.167	0.284	0.126	0.189
HMD	0.154	0.000	0.321	0.098	0.196	0.280	0.000	0.202	0.254	0.256	0.234	0.155	0.151
PRES	0.069	0.005	0.264	0.329	0.278	0.372	0.464	0.000	0.353	0.127	0.304	0.002	0.108
WSPD	0.328	0.297	0.423	0.069	0.188	0.276	0.437	0.211	0.000	0.004	0.083	0.083	0.206
TEMP	0.076	0.350	0.607	0.077	0.234	0.354	0.418	0.512	0.172	0.000	0.321	0.125	0.196
ACTV	0.172	0.258	0.140	0.026	0.234	0.413	0.280	0.004	0.145	0.131	0.000	0.052	0.150

CASES	0.000	0.001	0.005	0.272	0.504	0.092	0.144	0.484	0.130	0.582	0.877	0.000	0.957
DAYS	0.016	0.063	0.217	0.331	0.027	0.412	0.422	0.100	0.142	0.310	0.711	0.073	0.000
Cluster 3 – Post-peak phase													
	PM2.5	PM10	SO_2	CO	NO_2	O_3	HMD	PRES	WSPD	TEMP	ACTV	CASES	DAYS
PM2.5	0.000	0.389	0.050	0.686	0.341	0.479	0.662	0.278	0.021	0.223	0.173	0.090	0.143
PM10	0.833	0.000	0.475	0.755	0.517	0.080	0.405	0.002	0.082	0.221	0.251	0.034	0.154
SO_2	0.277	0.380	0.000	0.334	0.734	0.057	0.452	0.264	0.130	0.291	0.209	0.001	0.010
CO	0.037	0.243	0.207	0.000	0.208	0.058	0.592	0.041	0.270	0.162	0.001	0.161	0.225
NO_2	0.058	0.214	0.436	0.391	0.000	0.288	0.253	0.002	0.034	0.008	0.048	0.016	0.022
O_3	0.132	0.004	0.145	0.334	0.143	0.000	0.282	0.158	0.284	0.167	0.284	0.126	0.189
HMD	0.154	0.000	0.321	0.098	0.196	0.280	0.000	0.202	0.254	0.256	0.234	0.155	0.151
PRES	0.069	0.005	0.264	0.329	0.278	0.372	0.464	0.000	0.353	0.127	0.304	0.002	0.108
WSPD	0.328	0.297	0.423	0.069	0.188	0.276	0.437	0.211	0.000	0.004	0.083	0.083	0.206
TEMP	0.076	0.350	0.607	0.077	0.234	0.354	0.418	0.512	0.172	0.000	0.321	0.125	0.196
ACTV	0.172	0.258	0.140	0.026	0.234	0.413	0.280	0.004	0.145	0.131	0.000	0.052	0.150
CASES	0.000	0.001	0.005	0.272	0.504	0.092	0.144	0.484	0.130	0.582	0.877	0.000	0.957
DAYS	0.016	0.063	0.217	0.331	0.027	0.412	0.422	0.100	0.142	0.310	0.711	0.073	0.000

Appendix B Supplementary results for metformin transport in porous media

Table B. 1 Causal estimation results and backdoor variable sets

Treatment	Outcome	ATE for DML	ATE for LNR	Backdoor Variables
Particle Density	Hydraulic Conductivity	-0.8935	-0.6731	Porosity
Particle Density	Dispersivity	0.0303	0.0371	Degree of Saturation, Porosity, Distance
Particle Density	Adsorption Coefficient	0.5310	0.3335	Porosity
Particle Density	Type-1 Sorption Fraction	0.0211	-0.0090	Degree of Saturation, Porosity
Hydraulic Conductivity	Relative Velocity	0.0197	0.0339	Dispersivity, Type-1 Sorption Fraction, Particle Density, Horizontal, Distance, Adsorption Coefficient, Degree of Saturation, Flux, Porosity, Type-2 Sorption Reaction Rate, Concentration
Dispersivity	Relative Velocity	-0.0050	0.0005	Type-1 Sorption Fraction, Particle Density, Horizontal, Distance, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Flux, Porosity, Type-2 Sorption Reaction Rate, Concentration
Adsorption Coefficient	Relative Velocity	-1.1596	-1.0644	Dispersivity, Type-1 Sorption Fraction, Particle Density, Horizontal, Distance, Hydraulic Conductivity, Degree of Saturation, Flux, Porosity, Type-2 Sorption Reaction Rate, Concentration
Type-1 Sorption Fraction	Relative Velocity	-0.0400	-0.0423	Dispersivity, Particle Density, Horizontal, Distance, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Flux, Porosity, Type-2 Sorption Reaction Rate, Concentration
Porosity	Hydraulic Conductivity	0.5834	0.2094	Particle Density
Porosity	Dispersivity	-0.0259	0.0127	Particle Density, Degree of Saturation, Distance
Porosity	Adsorption Coefficient	-0.2073	-0.0712	Particle Density
Porosity	Type-1 Sorption Fraction	0.0392	0.0242	Particle Density, Degree of Saturation
Degree of Saturation	Dispersivity	0.0409	0.0483	Particle Density, Distance, Hydraulic Conductivity, Adsorption Coefficient, Porosity
Degree of Saturation	Type-1 Sorption Fraction	0.0177	0.0082	Particle Density, Porosity, Hydraulic Conductivity, Adsorption Coefficient
Degree of Saturation	Relative Velocity	0.2665	0.1625	Particle Density, Horizontal, Distance, Hydraulic Conductivity, Adsorption Coefficient, Flux, Porosity, Type-2 Sorption Reaction Rate, Concentration
Distance	Dispersivity	0.0014	0.0023	Type-1 Sorption Fraction, Particle Density, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Porosity
Distance	Relative Velocity	-0.0324	-0.0351	Type-1 Sorption Fraction, Particle Density, Horizontal, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Flux, Porosity, Type-2 Sorption Reaction Rate, Concentration
Concentration	Type-2 Sorption Reaction Rate		0.0001	
Concentration	Relative Velocity	-0.0005	0.0001	Dispersivity, Type-1 Sorption Fraction, Particle Density, Horizontal, Distance, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Flux, Porosity
Type-2 Sorption Reaction Rate	Relative Velocity	-0.1357	-0.1447	Dispersivity, Type-1 Sorption Fraction, Particle Density, Horizontal, Distance, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Flux, Porosity, Concentration
Ponded Water Depth	Relative Velocity	-0.0012	14.3241	Dispersivity, Type-1 Sorption Fraction, Particle Density, Horizontal, Distance, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Flux, Porosity, Type-2 Sorption Reaction Rate, Concentration
Flux	Relative Velocity	-0.0005	-0.0254	Dispersivity, Type-1 Sorption Fraction, Particle Density, Horizontal, Distance, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Ponded Water Depth, Porosity, Type-2 Sorption Reaction Rate, Concentration
Horizontal	Relative Velocity	0.0310	-0.0193	Dispersivity, Type-1 Sorption Fraction, Particle Density, Distance, Hydraulic Conductivity, Adsorption Coefficient, Degree of Saturation, Flux, Ponded Water Depth, Porosity, Type-2 Sorption Reaction Rate, Concentration

ATE: Average Treatment Effect; **DML:** CausalForestDML Estimator; **LNR:** Linear Estimator

Table B. 2 Causal refutation results

Treatment	Outcome	Refuter: RCC for DML	Refuter: UOC for DML	Refuter: PLB for DML	Refuter: RCC for LNR	Refuter: UOC for LNR	Refuter: PLB for LNR
Particle Density	Hydraulic Conductivity	-0.8913	-0.8192	0.0004	-0.6731	-0.6240	-0.0002
Particle Density	Dispersivity	0.0310	0.0190	-0.0009	0.0371	0.0387	0.0010
Particle Density	Adsorption Coefficient	0.5285	0.4471	0.0011	0.3335	0.2966	-0.0008
Particle Density	Type-1 Sorption Fraction	0.0239	0.0150	0.0001	-0.0090	-0.0034	-0.0012
Hydraulic Conductivity	Relative Velocity	0.0218	0.0301	-0.0002	0.0339	0.0290	0.0006
Dispersivity	Relative Velocity	-0.0057	-0.0069	0.0000	0.0005	-0.0006	-0.0006
Adsorption Coefficient	Relative Velocity	-1.1555	-0.9423	-0.0011	-1.0644	-0.9095	0.0007
Type-1 Sorption Fraction	Relative Velocity	-0.0396	-0.0395	0.0000	-0.0423	-0.0421	-0.0005
Porosity	Hydraulic Conductivity	0.5823	0.5429	-0.0006	0.2094	0.1756	-0.0006
Porosity	Dispersivity	-0.0302	-0.0362	0.0024	0.0127	0.0120	-0.0024
Porosity	Adsorption Coefficient	-0.2065	-0.1937	-0.0006	-0.0712	-0.0598	0.0001
Porosity	Type-1 Sorption Fraction	0.0257	0.0312	-0.0006	0.0242	0.0227	-0.0020
Degree of Saturation	Dispersivity	0.0468	0.0444	-0.0014	0.0483	0.0492	-0.0015
Degree of Saturation	Type-1 Sorption Fraction	0.0147	0.0234	0.0016	0.0082	0.0042	-0.0002
Degree of Saturation	Relative Velocity	0.2650	0.1149	-0.0001	0.1625	0.1518	0.0004
Distance	Dispersivity	0.0007	-0.0039	0.0000	0.0023	0.0017	0.0002
Distance	Relative Velocity	-0.0325	-0.0328	0.0000	-0.0351	-0.0356	0.0002
Concentration	Type-2 Sorption Reaction Rate				0.0001	-0.0024	-0.0024
Concentration	Relative Velocity	-0.0004	-0.0021	-0.0002	0.0001	-0.0006	-0.0002
Type-2 Sorption Reaction Rate	Relative Velocity	-0.1354	-0.1297	0.0002	-0.1447	-0.1360	-0.0011
Ponded Water Depth	Relative Velocity	-0.0010	-0.0300	0.0002	14.3153	-0.0222	-0.0003
Flux	Relative Velocity	-0.0010	-0.0124	0.0001	-0.0254	-0.0367	0.0001
Horizontal	Relative Velocity	0.0295	0.0285	0.0000	-0.0193	-0.0175	-0.0003

RCC: Add Random Common Cause Refuter; **UOC:** Add Unobserved Confounder; **PLB:** Placebo Refuter

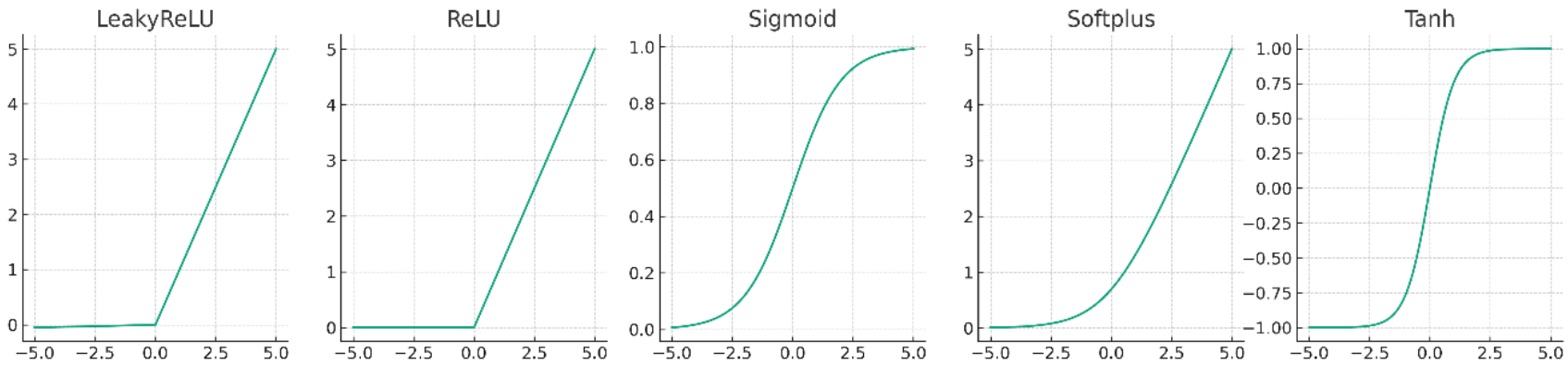


Figure B.1 Five activation functions discussed in the study. LeakyReLU allows a small, positive gradient when the unit is not active (i.e., when the input is negative). ReLU provides an output of 0 for all negative inputs and linear output for positive inputs. Sigmoid squashes its input into a range between 0 and 1 and is smooth and differentiable at every point. Softplus smoothly approximates the ReLU function. It is differentiable everywhere and its output is also in the range $(0, \infty)$. Tanh squashes its input into a range between –1 and 1. Like the sigmoid function, it's also smooth and differentiable at every point.

Appendix C Common activation functions in neural networks

Activation functions play a crucial role in neural network-based models, since appropriate activation functions are the key to desirable performance and the computational efficiency of an NN-based model. In a neural network, inputs are passed through layers of nodes (or "neurons"), each applying an activation function. The purpose of the activation function is to introduce non-linearity into the output of a neuron.

Here are the five activation functions discussed in this dissertation work:

(a) ReLU (Rectified Linear Unit):

$$\mathrm{f(x)} = \begin{cases} x & if\ x > 0 \\ 0 & otherwise \end{cases} \tag{C.1}$$

Pros: Computationally efficient and easy to implement. Non-zero gradients do not saturate, which is beneficial during the gradient descent process.

Cons: The 'dying ReLU' problem. For negative inputs, the gradient is zero, so once a neuron gets negative, it is unlikely to recover. This 'dead neuron' would then always output the same value.

(b) LeakyReLU:

$$\mathrm{f(x)} = \begin{cases} x & if\ x > 0 \\ ax & otherwise \end{cases} \tag{C.2}$$

Pros: Introduces a small slope to keep the updates alive, thus mitigating the 'dying ReLU' problem. Helps to keep some information flowing, even for negative input values.

Cons: The performance is not consistently better than ReLU in practice.

(c) Tanh:

$$\mathrm{f}(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{C.3}$$

Pros: Its output is zero-centered because its range is -1 to 1. This makes the model converge faster than when using the sigmoid function.

Cons: It still has the vanishing gradient problem for very large positive and negative values.

(d) Softplus:

$$f(x) = log(1 + e^{x}) \tag{C.4}$$

Pros: Softplus is smooth everywhere and its derivative (a version of the sigmoid function) is easy to compute, which could be beneficial in some cases.

Cons: It may suffer from numerical instability; for large inputs, the output of Softplus could be infinite.

(e) Sigmoid:

$$\mathrm{f}(x) = \frac{1}{1 + e^{-x}} \tag{C.5}$$

Pros: It is smooth and differentiable everywhere. Its output range between 0 and 1 is often useful in models like logistic regression, where the output can be interpreted as a probability.

Cons: It suffers from the vanishing gradient problem. Furthermore, its output is not zero-centered.

Appendix D Data Sources for Gross Domestic Product and Population

The data sources for Gross Domestic Product (GDP) per capita in Chapter 5 are detailed in this section. To accurately represent the economic status of smaller regions during the specified period, GDP data were meticulously gathered and consolidated, which varied in availability across different countries. Thus, the data collection procedures are outlined here for clarity. Unless otherwise stated, currency conversion was standardized using the exchange rates published by the Organisation for Economic Co-operation and Development (OECD) to ensure consistency and comparability across countries and regions.

Africa:

Egypt. The GDP data for the Asyut Governorate were retrieved from the Ministry of Planning and Economic Development's official website (https://mped.gov.eg/Governorate?lang=en). The population figures for the same region were obtained from the Egypt Central Agency for Public Mobilization and Statistics' Report on Births and Deaths (Central Agency for Public Mobilization and Statistics, 2020). The exchange rate data from CEIC were used to convert the Egyptian Pound (EGP) to US Dollars (USD).

Nigeria. The Lagos State government website (https://lagossdgandinvestment.com/glancelagos) provides the reported GDP per capita in USD for Lagos State.

South Africa. The Ekurhuleni City GDP figures were published by the Gauteng Provincial Treasury, while the GDP of the KwaZulu-Natal Province was obtained from

Statistics South Africa (Statistics South Africa, 2023). Additionally, population data for South African municipalities were sourced from local government publications, specifically “The Local Government Handbook: South Africa" (Yes! Media, 2023).

Uganda. Uganda’s national GDP per capita was sourced from the World Bank database, accessible at https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?locations=UG.

Asia

China. Annual economic reports, or annual government working reports for various Districts/Counties (level 4 administrative units under cities), were primarily referenced as the major data source. They are typically available on the respective government’s official website by year-end. These reports often detail the gross domestic product and the population of Usual Residents, allowing for the calculation of GDP per capita. Occasionally, GDP per capita figures are directly reported within these documents. When such reports were unavailable, statistical yearbooks published by the local governments were utilized. Currency conversion from Chinese Yuan (CNY) to current US Dollar (USD) values was based on the annual exchange rate published by the China Foreign Exchange Trade System.

Saudi Arabia. GDP data for Medinah were extracted from publications by the Saudi Arabian Ministry of Municipal and Rural Affairs (Saudi Arabian Ministry of Municipal and Rural Affairs, 2019). Population figures were sourced from the Population and Housing Census 2010 (https://portal.saudicensus.sa/portal/), which were then used to calculate the GDP per capita. The CEIC's exchange rates were applied to convert Saudi Riyals (SAR) to USD.

South Korea. District-level GDP, population, and GDP per capita data were obtained from the Korean Statistical Information Service (KOSIS, https://kosis.kr/eng/).

North America

Mexico. State-level (Level 1) GDP data were obtained from Mexico's Economic Information System (https://en.www.inegi.org.mx/app/indicadores/?tm=0). State population figures were sourced from the Statistical and Geographical Yearbook (National Institute of Statistics and Geography, 2015).

The United States of America. GDP per capita data for New York and Urbana-Champaign were acquired from FRED Economic Data, provided by the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/).

Europe

Generally, for European countries, including Belgium, Croatia, Germany, Italy, Sweden, and Switzerland, GDP per capita data were sourced uniformly from OECD statistics (https://stats.oecd.org/). Exchange rates were also obtained from the OECD, ensuring consistency in currency conversions.

Czech Republic. Regional GDP per capita data for the South Moravian Region can be found at the Czech Statistical Office (https://apl.czso.cz/pll/rocenka/rocenka.indexnu_reg?mylang=EN).

Faroe Islands. GDP per capita figures are published by Statistics Faroe Islands (https://hagstova.fo/en/economy/national-accounts/gdp-and-main-figures).

Hungary. The GDP per capita for Budapest was retrieved from the Hungarian Central Statistical Office, accessible at the following URL: https://www.ksh.hu/stadat_files/gdp/en/gdp0078.html.

Iceland. Statistics Iceland provides GDP per capita information (https://px.hagstofa.is/pxen/pxweb/en/Efnahagur/Efnahagur__thjodhagsreikningar__landsframl__1_landsframleidsla/THJ01401.px).

Slovakia. The GDP per capita for the Žilina Region was sourced from the Statistical Office of the Slovak Republic, available online here: https://datacube.statistics.sk/#!/view/en/VBD_SK_WIN/nu3002rr/v_nu3002rr_00_00_00_en.

Slovenia. GDP per capita data for Central Slovenia and the Lower Sava region were obtained from the Republic of Slovenia Statistical Office. The relevant data can be reviewed at: https://pxweb.stat.si/SiStatData/pxweb/en/Data/Data/0309250S.px/.

Romania. For Bucharest and Cluj, GDP and population data are available at the National Institute of Statistics, Romania (https://insse.ro/cms/).