Mathematical and Computational Modelling of biological Systems using Ordinary Differential Equations Framework and In-Silico Approaches

A Thesis Submitted

To

Sikkim University

In Partial Fulfilment of the Requirement for the

Degree of Doctor of Philosophy

By

Bikash Thakuri

Department of Mathematics

School of Physical Sciences

May 2023

6 माइल, सामदुर, तादोंग - 737102
गंगटोक, सिक्किम, भारत
फोन-03592-251212, 251415, 251656
टेलीफैक्स - 251067
वेबसाइट - www.cus.ac.in

6th Mile, Samdur, Tadong-737102
Gangtok, Sikkim, India
Ph. 03592-251212, 251415, 251656
Telefax : 251067
Website : www.cus.ac.in

सिक्किम विश्वविद्यालय
SIKKIM UNIVERSITY

(भारत के संसद के अधिनियम द्वारा वर्ष 2007 में स्थापित और नैक (एनएएसी) द्वारा वर्ष 2015 में प्रत्यायित केंद्रीय विश्वविद्यालय)
(A central university established by an Act of Parliament of India in 2007 and accredited by NAAC in 2015)

DECLARATION

I declare that the thesis entitled "**Mathematical and Computational Modelling of biological Systems using Ordinary Differential Equations Framework and In-Silico Approaches"** submitted by me for the award of **Doctor of Philosophy** in Mathematics of **Sikkim University** is my original work. The content of this thesis is based on the work which I have performed myself. This thesis has not been submitted for any degree to any other University. The content of this Ph.D. thesis has been subjected to the anti-plagiarism software (URKUND) and was found satisfactory.

Date 04/05/2023

(Bikash Thakuri)

Roll No.: 17PDMT01

Reg. No.: 17/Ph.D/MAT/01

Recommended that the thesis be placed before the Examiners for evaluation.

(Dr. Amit Chakraborty)
Head of Department
Department of Mathematics
School of Physical Sciences

अध्यक्ष/प्रभारी
Head/In-charge
गणित विभाग
Department of Mathematics
सिक्किम विश्वविद्यालय
Sikkim University

(Dr. Amit Chakraborty)
Ph.D. Supervisor
Department of Mathematics
School of Physical Sciences

6 माइल, सामदुर, तादोंग - 737102
गंगटोक, सिक्किम, भारत
फोन-03592-251212, 251415, 251656
टेलीफैक्स - 251067
वेबसाइट - www.cus.ac.in

6th Mile, Samdur, Tadong-737102
Gangtok, Sikkim, India
Ph. 03592-251212, 251415, 251656
Telefax : 251067
Website : www.cus.ac.in

सिक्किम विश्वविद्यालय
SIKKIM UNIVERSITY

(भारत के संसद के अधिनियम द्वारा वर्ष 2007 में स्थापित और नैक (एनएएसी) द्वारा वर्ष 2015 में प्रत्यायित केंद्रीय विश्वविद्यालय)
(A central university established by an Act of Parliament of India in 2007 and accredited by NAAC in 2015)

CERTIFICATE

This is to certify that the thesis entitled "**Mathematical and Computational Modelling of biological Systems using Ordinary Differential Equations Framework and In-Silico Approaches"** submitted to Sikkim University in partial fulfilment of the requirement for the degree of **Doctor of Philosophy** in Mathematics embodies the result of *bona fide* research work carried out by **Mr. Bikash Thakuri** under my guidance and supervision. No part of the thesis has been submitted for any other degree, diploma, associate-ship, fellowship.

All the assistance and help received during the investigation have been duly acknowledged by him.

Place: Gangtok, Sikkim
Date: 04/05/2023

(Dr. Amit Chakraborty)
Ph.D. Supervisor
Department of Mathematics
School of Physical Sciences

6 माइल, सामदुर, तादोंग - 737102
गंगटोक, सिक्किम, भारत
फोन-03592-251212, 251415, 251656
टेलीफैक्स - 251067
वेबसाइट - www.cus.ac.in

6th Mile, Samdur, Tadong-737102
Gangtok, Sikkim, India
Ph. 03592-251212, 251415, 251656
Telefax : 251067
Website : www.cus.ac.in

सिक्किम विश्वविद्यालय
SIKKIM UNIVERSITY

(भारत के संसद के अधिनियम द्वारा वर्ष 2007 में स्थापित और नैक (एनएएसी) द्वारा वर्ष 2015 में प्रत्यायित केंद्रीय विश्वविद्यालय)
(A central university established by an Act of Parliament of India in 2007 and accredited by NAAC in 2015)

Date: 04/05/2023

PLAGARISM CHECK REPORT

This is to certify that plagiarism check has been carried out for the following Ph.D. Thesis with the help of **URKUND Software** and the result is 1 % which is within the permissible limit (below 10% tolerance rate) as per the norms of Sikkim University.

"Mathematical and Computational Modelling of biological Systems using Ordinary Differential Equations Framework and In-Silico Approaches"

Submitted by **Mr. Bikash Thakuri** under the supervision of **Dr. Amit Chakraborty, Associate Professor, Department of Mathematics, School of Physical Sciences**, Sikkim University, Gangtok.

(Bikash Thakuri)

Signature of the Research Scholar

(Dr. Amit Chakraborty)

Signature by the Supervisor

4/5/2023

Vetted by Librarian

केन्द्रीय पुस्तकालयाध्यक्ष
Librarian
केन्द्रीय पुस्तकालय Central Library
सिक्किम विश्वविद्यालय
Sikkim University

Dedicated

To

My Parents,

Mentor

& My Loving Wife

Acknowledgments

I want to express my gratitude to a few of the many wonderful people I have met over this period. I want to start by sincerely thanking **Dr. Amit Chakraborty**, my Ph.D. supervisor and the man who launched my research career. His insightful knowledge and helpful advice gave me a solid platform for becoming an independent researcher. This thesis work would not have been possible without his mentorship, inspiration, and support from the very first day of my Ph.D. studies. I appreciate your advice and help very much.

My sincere and heartfelt gratitude to **Dr. Roshan Thoudam Singh** and Dr. **Bipul Pal**, my previous supervisors. Along with this, I would like to extend my sincere gratitude **to Dr. Namita Behera, Mrs. Rinkila Bhutia**, and the entire faculty from the Department of Mathematics, Sikkim University for their assistance during the time of my Ph.D. work. My sincere gratitude to **Mr. Yogen Ghatani** and **Mr. Naseem Zoha Ansari**, members of my whole team, for their ongoing support and recommendations for my study. During lab discussions, I always appreciated your ideas and recommendations, and when you are kind outside of the lab, it makes me pleased, especially during trying moments.

I would like to extend my sincere gratitude and debt of gratitude to our research partners **Professor Ce Wang** from **Southeast University in Nanjiang, People's Republic of China**, **Dr. Adnan Sljoka** from **the RIKEN Centre for Advanced Intelligence Project in Tokyo, Japan**, **Professor Swarup Roy** from the **Department of Computer Applications, Sikkim University**, **Dr. Nitish Mondal** from the **Department of Anthropology in the School of Human Sciences at**

Sikkim University, Dr. Amit Kumar Roy Department of Mathematics, Sikkim University, **Dr. Jayanta Kumar Das, Postdoc (John Hopkins University, USA), Research Fellow at the National Institutes of Health**, and **Prof. Gennady N. Chuev; The Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Moscow, Russia** for their active collaboration, visiting my lab, sharing ideas, providing comments and suggestions related to my research problem.

I would like to express my deep sense of gratitude to our **Hon'ble Vice Chancellor Prof. Avinash Khare, Sikkim University** for all kinds of support gifted to me.

For their unwavering love, support, prayers, and various forms of encouragement**, Lt. Prem Thakuri (Father), Mrs. Kumari Thakuri (Mother), Mr. Mahesh Thakuri (Brother),** and **Mrs. Banita Rai (Wife)** are much appreciated. In closing, I praise God for his unfailing grace, love, and flawless righteousness during my whole research career.

(Bikash Thakuri)

Abstract

The study of biological systems has benefited greatly from mathematical modelling since it us to replicate complicated biological processes and better understand the underlying mechanisms. ODEs, which characterise the rate of change of a system's variables over time and are especially helpful for systems susceptible to continuous changes, are a key tool for this kind of study. We may examine how various factors change over time and forecast system behaviour by modelling biological processes as mathematical equations. These models may be used to create fresh experiments and test theories. In silico research, which makes use of computer simulations to analyse the behaviour of biological systems under diverse conditions, is one crucial use of mathematical modelling in biology. Advancing our knowledge of biological systems and creating fresh approaches to curing disease and enhancing human health depend on the integration of mathematical models and experimental data.

In different chapters of the present study, various types of data are analysed using suitable sophisticated mathematical tools and techniques while aiming for multiple plausible solutions to a variety of biological and ecological problems. For instance, adverse feedback effects brought on by AT2R-Ang II binding reactions serve as reliable controls on mean-arterial variance, which, if dysregulated, might result in either hypertension or hypotension. Spatial temporal data analysis of PAHs concentration of South China Sea (SCS) revealed significant seasonal fluctuation in PAH levels, and the most significant environmental variables impacting the seasonal heterogeneity and the geographical distributions of PAHs in the surface sea waters are considered to be anthropogenic activities, land- and ocean-based emissions, surface runoff, and open seawater dilution. Another chapter analysed the ecological risk

evaluation of OCPs seasonal and phase-petitioning effects in SCS and ECS, which are important source-sink zones. Overall, the research emphasizes the significance of taking spatiotemporal variation into account when assessing ecological risk. In addition, time series analysis of data on malaria incidence in the Asia-Pacific area and Africa revealed significant intra- and inter-regional variations, with almost exclusively decadal declining patterns in the Asia-Pacific region and a somewhat mixed tendency in Africa. This study illustrates the relationship between intra- and inter-regional variations in malaria incidence and environmental conditions at broad geographical scales. Furthermore, understanding the significance of mutations in the SARS-CoV-2 spike protein, including E484K, K417N, L452Q, L452R, N501Y, and T478K, can aid in targeted control measures, laboratory characterization, and therapeutic efforts. These mutations are part of an allosteric network that affects interactions between the spike protein and human receptor ACE2, leading to higher transmissibility and infectivity. Compensatory mutations in the N-terminal domain (NTD) are also involved in this network, and mutations in the RBD increase interactions with ACE2 to varying extents, depending on their allosteric connections with compensatory mutation clusters in the NTD.

Table of Contents

Acknowledgments.....i

Abstract..... iii

Table of Contents.....v

List of Figures.....xi

List of Tables.....xiv

1. Introduction.....1

1.1 Motivation.....1

1.1.1 High-throughput data.....2

1.1.2 Heterogenous data.....2

1.1.3 Homogenous data.....3

1.1.4 Time series data.....3

1.1.5 Spatiotemporal data.....4

1.1.6 Synthetic data.....5

1.2 Recent development and importance.....6

1.2.1 National importance.....6

1.2.2 International importance.....7

1.3 Literature review of data types.....7

1.4 Research objectives.....10

1.5 Layout of the thesis 11

2. Synthetic data modelling.....13

2.1 Introduction 14

2.1.1 Angiotensin II subtype-1 receptor (AT1R) and Angiotensin II subtype-1 receptor (AT1R).....15

2.1.2 Vasodilator and hypotensive effects of AT2R.....16

2.2 Methods.....17

2.2.1 Mathematical model of cRAS.....19

2.2.2 Canonical model for circulating-RAS 20

2.3 Numerical solutions of non- linear ODEs.....24

2.4 Non-linear least square method.....25

2.5 *lsqnonlin* function.....26

2.6 Kolmogorov-Smirnov test.....26

2.7 Local sensitivity analysis 27

2.8 Synthetic data generation and in-silico Experiment.....28

2.9 Results.....32

2.10 Summary 41

3. Spatiotemporal data modelling and analysis-I.....42

3.1 Introduction.....43

3.1.1 PAH species.....45

3.2 Methods and materials 46

3.2.1 Study area and sample collection.....46

3.2.2 Data analysis 47

3.3 Results 52

3.3.1 Spatial distribution and PAH phase composition 52

3.3.2 Correlations and seasonal heterogeneity of PAH 54

3.3.3 Effects of PAH phase partitioning and source location 56

3.3.4 Evaluation of the ecological risks posed by dissolved and particulate PAHs 61

3.4 Summary 63

4. Spatiotemporal data modelling and analysis-II.....65

4.1 Introduction.....66

4.2 OCPs species.....68

4.3 Methods and materials 68

4.3.1 Sample collection and the study region 68

4.4 Data analysis 72

4.4.1 K-means clustering 72

4.4.2 Principal Component Analysis (PCA) 73

4.4.3 Hausdorff Distance Measure (HDM).....74

4.4.4 Mixture risk model (MRM) 75

4.5 Results 76

4.5.1 Composition and geographic spread of the OCP phase 76

4.5.2 Different prevalent OCP species have distinct impacts. 82

4.5.3 Evaluation of the spatiotemporal ecological risk 84

4.6 Discussion 88

4.6.1 OCPs' spatiotemporal heterogeneity 88

4.6.2 OCP sources in the SCS and ECS 90

4.6.3 Maps of ecological risk 101

4.7 Summary 104

5. Time series data modelling and analysis 105

5.1 Introduction 105

5.2 Methods 107

5.2.1 Data collection 107

5.2.2 Generalized linear model and mixed effects model 111

5.2.3 Agglomerative clustering 112

5.3 Results 113

5.3.1 Malaria prevalence and climatic variables vary intra- and inter-regionally. 113

5.3.2 Modelling the relationships between malaria prevalence and climatic influences 121

5.4 Discussion 134

5.5 Summary 138

6. Protein data analysis and applications 140

6.1 Introduction 141

6.2 Results and Discussion 146

6.2.1 Network of compensatory mutations in SARS-CoV-2 spike types.....146

6.2.2 Long-range dynamic allostery is the driving force behind mutational spots in the receptor-binding domain (RBD). 152

6.2.3 Spike protein associations with ACE2 are modified by mutations through a dynamic allosteric network. 160

6.3 Methods 166

6.3.1 Detecting mutational sites that are susceptible across SARS-CoV-2 spike types 166

6.3.2 The extraction of cohesive mutant sites. 167

6.3.3 Residues between RBD and NTD undergo chemical shift changes and allosteric coupling 168

6.3.4 Utilizing 3D protein structures for interface analysis 172

6.4 Summary 174

Future direction 176

Bibliography 177

Appendix A 195

List of program Codes 195

MATLAB code: Parameters estimation.....195

MATLAB code: Ordinary differential equation 198

MATLAB code: Parameter Estimation 200

MATLAB code: Solution of differential equation.....201

MATLAB code: Transient Analysis.....203

MATLAB code: RtoODE Function.....205

Appendix B 206

List of Publications 206

Appendix C 208

List of Seminar/Conferences/Workshops attained.....208

List of Figures

Figure 2.1 Schematic diagram of AT2R mediated feedback effects on Mean Arterial Pressure Regulation (MAP) 13

Figure 2.2 Block diagram of the renin-angiotensin (cRAS) system 18

Figure 2.3 Insilico experiment flow chart 30

Figure 2.4 Transient dynamic of [Renin and [ANGII] 34

Figure 2.5 Q-Q plots showing deviations of [Renin] and [ANG II] 35

Figure 2.6 A linear steady-state relationship between [Renin] and [ANGII] 39

Figure 3.1 Graphical Abstract 42

Figure 3.2 Sampling location along the South and East China Seas 48

Figure 3.3 Spatial distribution of 16 PAH identified by U.S. EPA 54

Figure 3.4 Correlation and distance matrix of all PAH species 56

Figure 3.5 Non-Metric multidimensional scaling (NMDS) analysis 57

Figure 3.6 NMDS Season-Wise 58

Figure 3.7 Risk quotient (RQ) variation across the four Seasons 62

Figure 4.1Graphical Abstract 65

Figure 4.2 Sampling location along the South and East China Seas 71

Figure 4.3 K-means clustering under L1 distance matric 79

Figure 4.4 Hausdorff Distance Map 81

Figure 4.5 Principal component Analysis 82

Figure 4.6 Mixture risk Model estimated ecological risk posed by major OCP classes 86

Figure 4.7 Spatiotemporal ecological risk maps 104

Figure 5.1 Two most Malaria -affected global areas, Asia Pacific and Africa 115-118
Figure 5.2 Distribution of Malaria incidence 119
Figure 5.3 Cook's distance 126
Figure 5.4 The normal probability plot of residual results 128
Figure 5.5 The forest Plot showing fixed effects 129
Figure 5.6 The forest plot showing random effects 130
Figure 5.7 Scatter graphs of GLM-G estimates of temperature and precipitation factors 131
Figure 5.8 Agglomerative clustering 132
Figure 5.9 Agglomerative clustering 133
Figure 6.1 A probability scheme for identifying compensatory changes based on amino-acid (AA) sequences 147
Figure 6.2 The "AA compensatory mutation network (CMN)" for the SARS-CoV-2 spike protein 150
Figure 6.3 The mapping of compensating mutation regions onto the protomer of the spike(S) protein in three dimensions using RBD-up form 151
Figure 6.4 The allosteric interaction between VOI/VOC- specified RBD mutant site I and j in C1 C2 and C3 154
Figure 6.5 The regularity with which each residue is impacted by a change as a result of VOI/VOC- specified RBD mutations. 158
Figure 6.6 The chemical shift projection study demonstrating the consequences of particular RBD mutations for VOI/VOC 159

Figure 6.7 Impact of SARS-CoV-2 RBD and ACE2 interactions with various RBD angles and VOI/VOC-specified RBD mutations. 163

Figure 6.8 Multimodal degree distribution 164

List of Tables

Table 2.1 List of cRAS model parameters, together with definitions and references to standard values for each.	23
Table 2.2 Kolmogorov-Smirnov (KS) test results of significant differences among the three MAP condition	36
Table 2.3 Negative feedback control parameters in the RAS model in different MAP condition	38
Table 2.4 Local sensitivity analysis of AT2R-mediated feedback control parameters	40
Table 3.1 Season wise stress values of the NMD	58
Table 3.2 Principal Components Analysis (PCA) Result	60
Table 4.1 Basic chemical properties of 11 OCPs grouped into three primary classes	70
Table 4.2 Seasonal and phase-wise occurrences and distribution of major OCPs	78
Table 4.3 Ecological mixture risk assessment across seasons and phase in South China Sea and East China Sea	87
Table 4.4 Source Analysis of Particulate OCPs	93-96
Table 4.5 Source Analysis of Dissolved OCPs	97-102
Table 5.1 Model fit statistic	122
Table 5.2 GLM estimates of association coefficients for temperature and precipitation	123-124
Table 5.3 GLME estimates of random effect coefficients (Alpha 0.05)	126-127

Table 6.1 SARS-CoV-2 variant of concern (VOC) and variant of interest (VOI) 145

Table 6.2 The percentage of residues in the compensatory mutation regions C1,C2 and C3 with the total chemical shift-based association $|r| \geq 0.7$ and an allosteric link to the RBD site 155

Table 6.3 Interface characteristics of the S-RBD and ACE2 interactions under Six RBD variants with various RBD orientations that are unique to VOI/VOC 165

Chapter 1

1. Introduction

Mathematical modelling is increasingly important in the study of biological systems, allowing us to simulate complex biological phenomena and gain insight into the mechanisms that govern them. ODEs are a primary tool for this type of research, describing the rate of change of a system's variables over time and being particularly useful for systems subject to continuous changes.

By representing biological processes as mathematical equations, researchers can explore how different variables change over time and predict system behaviour under different conditions. These models can also be used to test hypotheses and design new experiments. An essential use of mathematical modelling in biology is in silico research. These studies make use of computer simulations to examine the behaviour of biological systems under various circumstances. However, mathematical models have limitations, including oversimplification and dependence on experimental data accuracy.

Nonetheless, combining mathematical models with experimental data is essential for advancing our understanding of biological systems and developing new strategies for treating disease and improving human health.

1.1 Motivation

Present study is essentially being motivated by the digital invitations taken place in last two decades, leading to enormous opportunities to test and validate our classic mathematical ideas and models while dealing with some core real life problems.

Following are the different types of data currently available in different public domains that can mostly be utilized with the use of suitable sophisticated mathematical tools and techniques while aiming for multiple plausible solutions of a variety of biological and ecological problems.

1.1.1 High-throughput data

High-throughput data is a term used to describe large amounts of data that are generated rapidly. In the fields of transcriptomics, proteomics, and genomics, where extensive study may rapidly produce enormous amounts of data, this phrase is frequently used. High-throughput data are produced using a variety of techniques, including protein interaction analysis, gene expression analysis, and DNA sequencing. One benefit of high-throughput data is the ability to look at multiple samples at once, which leads to a more comprehensive understanding of biological processes. This information can also be used to develop new diagnostic tools, find innovative drug targets, and broaden our understanding of illness. However, due to the size and intricacy of the actual data sets, high-throughput data can be difficult to handle and analyse. This requires the use of complex computational tools and methods, as well as storage choices to manage the enormous data volumes, to assess and comprehend the data.

1.1.2 Heterogenous data

Heterogeneous data is data that consists of various kinds, forms, formats, and sources. It is a common aspect of large data sets and can be challenging to effectively handle and assess. One of the main benefits of heterogeneous data is the amount of knowledge it offers. By combining data from various sources, we may gain a deeper

grasp of the subject. This could lead to wiser decision-making and more well-informed tactics. However, there are several challenges that come with diverse data. Integrating different types of data can be difficult because they may have contradictory patterns and formats. This calls for the use of specialised tools and techniques, such as data normalisation and mapping, to ensure that the data can be used effectively. Data diversity may also make research more challenging. It can be difficult to determine which types of data to use for a particular analytic activity because the quality, relevance, and reliability of different data types can change.

1.1.3 Homogenous data

Homogeneous data are those that share the same form, kind, or character. In other words, homogeneous data is information that is composed of similar parts and can be easily managed, processed, and analysed as a unique unit. Because it is easy and homogeneous, homogeneous data is important because it is simple to change, analyse, and gain deep insights. When data is homogenous, analysis can be performed completely without the need for pre-processing, uniformization, or other sanitization steps. While saving time and money, it also ensures the accuracy of the research. Homogeneous data, in general, is a crucial part of data administration and analysis, necessary for making data-driven choices.

1.1.4 Time series data

A time series is a collection of data that has been gathered over time, commonly at regular periods. Time series data can be found from a variety of sources, including financial markets, weather patterns, and manufacturing processes. Unlike cross-sectional data, which are snapshots of data collected at a specific point in time, the

data is distinct from other kinds of data because it has a temporal dimension. Time series data are commonly used in statistical analysis and modelling, such as trend analysis, forecasting, and anomaly detection. Time series models can be based on statistics or machine learning and can predict future values by taking into consideration patterns, periodicity, and outside occurrences. It can be challenging to work with time series data because of non-stationarity, missing numbers, and the frequency of outliers. Specific techniques and processes are required in order to effectively assess and model the data.

1.1.5 Spatiotemporal data

Spatiotemporal data is a general term for information that includes both a geographical location and a time component. This kind of information is becoming more and more prevalent in several fields, including geology, earth sciences, municipal planning, and meteorology. One of the primary advantages of spatiotemporal data is its ability to monitor changes over time and space. When compared to data that only has a spatial or temporal component, this can provide crucial insights into connections, patterns, and trends that are otherwise difficult to detect. A spatiotemporal collection of weather trends could be used, for instance, to study climate change and its impacts on different locations. Another advantage of spatiotemporal data is the ability to create interactive visualisations that can help illustrate intricate patterns and connections. However, managing spatiotemporal data can be challenging because it frequently includes large, complex datasets that call for specialised tools and approaches for processing, analysis, and display. The spatial and temporal elements of the data may also introduce new challenges, such as missing data, measurement errors, and spatial and temporal dependency. Spatiotemporal data,

in summary, is a useful instrument for understanding complex patterns and relationships across space and time. Although handling this type of data has some challenges, the benefits it provides make it an invaluable resource in many various industries.

1.1.6 Synthetic data

Synthetic data, also known as artificial or produced data, is a type of data that is made using computer programmes as opposed to being collected from trusted sources. Synthetic data is commonly used in a wide range of uses, including developing and testing machine learning models, enhancing data, and protecting and privacy. One of the primary benefits of synthetic data is the ability to quickly and easily create huge amounts of data. This can be particularly useful when getting real-world data is challenging or unattainable. Additionally, synthetic data can be modified to fit specific needs and requirements, for instance, by mimicking certain groups or trends found in real data. In general, synthetic data can be used for several tasks. It is important to remember, though, that artificial data might not truly represent real-world data and might not possess certain characteristics or complexity. Therefore, it is essential to carefully evaluate the quality and applicability of generated data for a given use case before using it. To sum up, synthetic data is a useful instrument with the ability to be very helpful in a range of uses. It is essential to use it carefully and with a clear understanding of its limitations if you want to ensure that the results are reliable and precise.

1.2 Recent development and importance

1.2.1 National importance

Understanding the significance of various kinds of data is essential in the rapidly evolving area of data science. All types of data, including heterogeneous and uniform data, spatiotemporal data, high-throughput data, synthetic data, and protein data, are crucial in a variety of industries, including agriculture, biotechnology, medicine, and many others. These kinds of information are crucial for innovation, study, and growth in India.

A mix of organized and unstructured data from various sources and forms is known as heterogeneous data. Although difficult to incorporate and evaluate, it is crucial to the Indian healthcare sector. Healthcare workers can develop a thorough grasp of a patient's health state and make well-informed therapy choices by integrating patient records, lab findings, and medical images. The spatial and temporal components of spatiotemporal data are essential to India's agricultural, healthcare, and environmental sectors. Data produced rapidly and in large quantities using a variety of methods, such as next-generation sequencing, is referred to as high-throughput data. This information is essential for locating genetic variations linked to illnesses, creating novel therapies, and carrying out extensive clinical studies in India's biotechnology and healthcare sectors. Artificially produced data, or synthetic data, is used to mimic real-world statistics. Large-scale data training of algorithms is becoming increasingly crucial in India's machine learning and artificial intelligence sectors. Machine learning models are more accurate when more varied databases are created, bias is reduced, and synthetic data is used. Analysis of proteins, essential components in living things, yields statistics on proteins. As it is used to create novel

medications and treatments, this information is crucial to the science and healthcare sectors in India. To find novel drug targets, comprehend disease processes, and create more potent therapies, researchers evaluate protein data. In conclusion, comprehension of the significance of various kinds of data is crucial for research, development, and invention across a range of industries, including agriculture, nanotechnology, and many others. Heterogeneous data, uniform data, spatiotemporal data, high-throughput data, synthetic data, and protein data are all essential for advancing technology and raising everyone's standard of living in India, where these sectors are expanding quickly.

1.2.2 International importance

Heterogenous data, homogenous data, spatiotemporal data, high throughput data, synthetic data, and protein data, all of these kinds of data are crucial for advancing science innovation and study on a worldwide scale, as well as for tackling issues like healthcare, climate change, and sustainable development. These kinds of data must be accessible and available if researchers and organizations from around the globe are to collaborate and share knowledge.

1.3 Literature review of data types

Several methods, including DNA sequencing, microarrays, mass spectrometry, and imaging, are used to provide high throughput data. High throughput data analysis requires specialised software tools and processing resources in order to be effective. Large datasets may be analysed to find patterns and trends using machine learning and artificial intelligence approaches. Applications for high throughput data are numerous, ranging from fundamental research to clinical trials. High throughput data

is a fast developing industry, with new technology and applications appearing often [1, 2].

The integration and analysis of data from many sources, including clinical, genomic, and imaging data, is referred to as heterogeneous data. Data fusion and machine learning are two examples of specialised software tools and methodologies that may be used to integrate and analyse heterogeneous data effectively. Numerous fields, such as illness diagnosis, drug development, and personalised medicine, make use of heterogeneous data analysis. Concerns about data privacy, data quality, and the necessity for cross-disciplinary collaboration are among the difficulties in heterogeneous data analysis[3].

Data that is homogeneous in nature and has the same structure and format is referred to as homogeneous data. It is frequently employed in applications for machine learning and statistical analysis. When compared to heterogeneous data, which has different forms and structures, homogeneous data is frequently simpler to analyse and handle. Numerical, textual, and visual data are typical types of homogenous data. Numerous industries, including healthcare, banking, and social research, employ homogeneous data. However, there are drawbacks to using homogenous data as well, including the risk of oversimplification and the requirement for high-quality data[4].

A group of data points that are accumulated through time are referred to as time series data. In several disciplines, including finance, economics, environmental research, and healthcare, time series data is extensively employed. Time series data processing calls for specialised methods including Fourier analysis, exponential

smoothing, and autoregressive integrated moving average (ARIMA). Time series data analysis has also employed machine learning methods like neural networks and support vector machines (SVMs). Forecasting future values, spotting anomalies, and identifying patterns are some applications of time series analysis[5].

Data that is gathered over both place and time is referred to as spatiotemporal data. The spread of illness, traffic patterns, and weather patterns are a few examples of spatiotemporal data. Spatiotemporal data analysis poses obstacles in terms of data management, quality, and the requirement for specialised analytical methods. Space-time clustering, spatial interpolation, and geographic information systems (GIS) are a few of the analytic methods utilised in spatiotemporal data analysis. Urban planning, illness surveillance, and climate modelling are a few examples of spatiotemporal data analytic applications[6, 7].

Data that is artificially produced as opposed to being gathered from authentic sources is referred to as synthetic data. To address privacy issues and to augment sparse or partial data, synthetic data can be employed. Generative adversarial networks (GANs), variational autoencoders (VAEs), and differential privacy are methods for producing synthetic data. Numerous industries, including healthcare, banking, and social research, use synthetic data. The necessity for validation against actual data and the possibility of overfitting are obstacles in the usage of synthetic data[8].

1.4 Research objectives

- **Objective # 1: To develop in-silico approaches to examine some system-regulatory mechanisms.**

  Developing simulations or models using computers is known as an in-silico method. Examining "system-regulatory mechanisms", which are the procedures that control how biological systems behave, is possible using these in-silico techniques. Because they enable fast and low-cost experimentation without the need for pricey physical experiments, in-silico methods for studying system-regulatory processes can be particularly helpful. Additionally, compared to conventional laboratory studies, computer-based simulations can offer a more thorough understanding of intricate biological systems.

- **Objective # 2: To develop some mathematical models to fit the real existing data with predicted values of certain biological systems**

  The process of developing mathematical models that can precisely forecast the behaviour of a biological system based on existing data is crucial. It usually starts by collecting data from real experiments or observations of the biological system before developing these models. Then this information is used to develop mathematical models or equations that explain the system's fundamental workings. In general, the creation of mathematical models that accurately reflect real-world data is an essential tool for understanding complex biological systems. It enables us to try and improve their hypotheses and can offer a more precise and thorough understanding of how these systems work.

- **Objective # 3: To probe the link between system properties, functional and dynamics across scales**

  Investigating the connections between a system's characteristics, how it functions, and how it evolves over time at various levels of structure is necessary to understand the association between system properties, functional, and dynamics across scales. The dynamics of the system as a whole start to matter at the biggest dimensions. For instance, variables like temperature change, species relationships, and human activities can have an impact on the dynamics of an ecosystem. We can better comprehend how complex systems function and create management plans for them by investigating these connections.

1.5 Layout of the thesis

The Ph.D. thesis is composed of six chapters that collectively explain and analyse the research problem, methodology, and findings of my study.

Chapter 1: This chapter includes a summary of the research's problems and findings.

Chapter 2: This chapter introduces the production of synthetic data using an in-silico method in the well-known Renin Angiotensin system, where we have found that the AT2R-Ang II binding reactions' negative feedback effects serve as reliable limits on mean-arterial variation. Depending on the degree of deviation, this control's instability may cause either hypertension or hypotension**.**

Chapter 3: A study of the spatiotemporal data analysis of the PAH concentration in the South China Sea (SCS) and East China Sea (ECS) has performed in this chapter.

PAHs concentration were collected for four seasons in two phases, particularly dissolved and particulate phase.

Chapter 4: In this chapter, a study of a spatiotemporal data analysis of OCPs distribution in the South China Sea (SCS) and East China Sea (ECS) was performed. Seasonal and phase-partitioning effects were explicitly considered when assessing ecological risk in this chapter. Two phases, specifically the dissolved and particulate phases, were used to gather OCP concentrations over the course of four seasons.

Chapter 5: In this chapter a "Heterogenous Time Series" data analysis of Malaria Incidence in two malaria endemic region, Asia-Pacific region and Africa was conducted. The study was conducted to identify intra-and inter regional differences in malaria incidence, if it exists in both the regions. Furthermore, the association of malaria incidences with temperature and precipitation was also studied in this chapter.

Chapter 6: In this chapter a highly transmissible SARS-CoV-2 variants with multiple spike mutations was studied which poses significant challenges in controlling the COVID-19. The study revealed that each RBD mutation site is engaged in an inter-domain allosteric network involving distant domain mutation sites, which affects interactions with the human receptor ACE2

The thesis ends with some limitations and suggestions for the future.

"Synthetic data is the gasoline of AI development". … Andrew Ng

Chapter 2

2. Synthetic data modelling

Graphical View

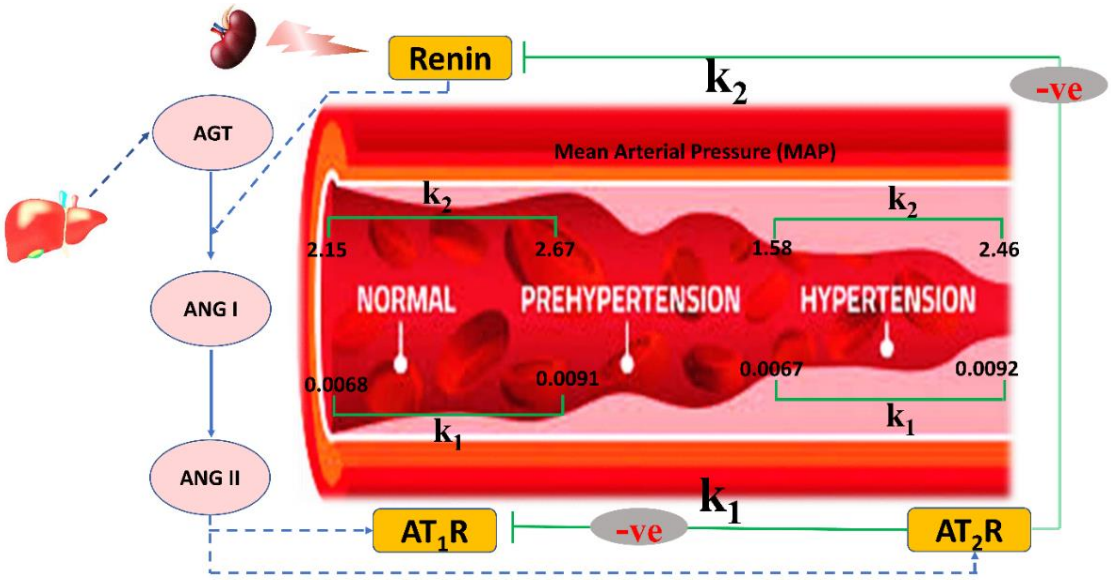


Figure 2.1 Schematic diagram of AT_2R mediated feedback effects on Mean Arterial Pressure Regulation (MAP)

Thakuri B, Das JK, Roy A.K., Chakraborty A.: Mean-arterial pressure maintenance under feedback controls over the circulating renin-angiotensin systems**. (Communicated)**

2.1 Introduction

The classical Renin Angiotensin System (RAS) model considers the canonical axis of the RAS systems that measure the harmful effects of Low Blood Pressure (LBP) that are mediated by the AT1R. The conventional belief that the RAS can only have negative impacts on the cardiovascular systems has been refuted by the recent finding of non-canonical RAS components. Angiotensins 1-7, 1-9, type 2 angiotensin II receptor, angiotensin-converting enzyme, and the proto-oncogene Mas receptors are among the non-canonical axis of the RAS that have been identified recently[9]. The negative consequences of classic RAS are offset by each component. Additionally, these non-canonical elements have been identified as prospective therapeutic targets because of their demonstrated essential roles in the pathophysiology and progression of several CVD [10-13]. Recent research has identified AT2R as the key factor mediating the majority of the effector Ang II's counter-regulatory effects among all the counter-regulating RAS axis [14, 15]. Mice missing or overexpressing this receptor show AT2R activation and associated vasodilatory effects [16]. While transgenic overexpression of the AT2R in vascular smooth muscle cells has demonstrated to prevent vasoconstriction, AT2R under-expression led to higher blood pressure levels. The therapeutic study of AT2R is advanced by the recent discovery of selective AT2R agonists CGP42112A, which has been demonstrated to lower blood pressure levels compared to untreated rats [17].

The main tenet of cardiovascular physiology is the Renin-Angiotensin System (RAS), a complex multi-organ endocrine system made up of several peptides and pathways [18]. The RAS works as a cascade and is frequently activated in response to low blood pressure (LBP) or damage. When the kidney's juxtaglomerular cells pump

the enzyme renin into the bloodstream, it begins acting on the liver-produced target protein angiotensinogen (AGT), which is continually present in plasma [19]. The AGT is broken down by renin into the inactive peptide angiotensin I. The angiotensin-converting enzyme (ACE) then transforms this peptide into angiotensin II (Ang II), a pleiotropic octapeptide, is the primary effector of the RAS system, which triggers a number of physiological reactions vital to cardiovascular disorders (CVD), such as hypertension and heart failure. This group of actions mostly functions by binding to two primary types of G-protein-coupled receptors with high affinity.[20], angiotensin II subtype-1 receptor (AT1R) and angiotensin II subtype-2 receptor (AT2R) (**Figure 2.2**). The AT1R mediates most of the well-known effects of RAS, which are typically harmful. However, little is known about the biochemical and functional consequences mediated by AT2R.

2.1.1 Angiotensin II subtype-1 receptor (AT1R) and Angiotensin II subtype-1 receptor (AT1R)

Only 34% of the sequence homology between the AT1R and the angiotensin II subtype-2 receptor (AT2R), which has a 363 amino acid molecular weight of 41,220 Da, has been preserved [21], despite the fact that they both have a high affinity for Ang II and are members of the G protein-coupled receptor family. However, AT2R predominates over AT1R in some specific areas of the uterus, ovary, adrenal medulla, and in some region of the brain. In most adult tissues of the kidney, adrenal cortex, and heart, AT2R expressions are substantially lower than AT1R [22-24]. There is increasing evidence that AT2R inhibits Ang II's vasoconstrictor effect mediated by AT1R [9]. Through the stimulation of a cascade made up of BK, NO, and cGMP, such AT2R-mediated effects culminate in vasodilation[25]. The production of NO

and cGMP is enhanced by AT2Rs either directly or indirectly through the stimulation of increased BK production, which is then mediated by BK B2 receptors[26]. To improve the treatment targets for CVDs, it is necessary to better understand how other complicated varied interactions, whether direct or indirect, mediated by the AT2R affected blood pressure control.

2.1.2 Vasodilator and hypotensive effects of AT2R

It is noticed that the vasodilator and hypotensive effects of AT2R are prolonged acutely for a long time and are not linked to desensitisation [24]. As a result, it has become an appropriate option for a therapeutic target in hypertension [27]. Indeed, prolonged Ang (AT1R) receptor blocker (ARB) administration encourages the AT2R activity with stronger vasodilator responses to Ang II, as shown by a recent in vitro Ang II-study of diabetic, hypertensive human individuals [27]. Similar results were observed in hypertensive rats following pharmacological AT2R activation in the presence of ARBs using Compound 21, a highly selective non-peptide agonist [28]. A completely distinct recent investigation in obese Zucker rats found that PD-123319's chronic AT2R antagonism increased mean arterial pressure (MAP) by 13 mm Hg and renin expression in the kidney cortex by three times [29]. In order to properly treat hypertension and hypotension, these results suggest the therapeutic benefit of examining various RAS interactions mediated by AT2R. The goal is to mimic AT2R's inhibitory feedback effects on AT1R and renin and investigate how these effects affect blood pressure control. It has been noticed that the AT2R-mediated inhibition of renin activity, which is frequently disregarded, has more drastic effects on MAP regulation than its opposing effects on AT1R. Additionally, it demonstrates

that under situations of hypertension and hypotension, these strict feedback controls of the MAP are relaxed.

2.2 Methods

The circulating RAS (cRAS) is made up of the precursor angiotensinogen, two important enzymes (renin and angiotensin-converting enzyme, ACE), and their bioactive by-products, angiotensin-I and II together with its receptors AT1R and AT2R. Multiple cascading effects are produced by the major effector component Ang II, principally through the AT1R and AT2R receptors (**Figure 2.2**). When Ang II binds to AT1R, a number of cytoplasmic signalling pathways are made possible, including those that control the contraction of vascular smooth muscle cells by activating myosin light chain kinase or inhibiting myosin light chain phosphatase [30-32]. Through the activation of the vasodilator cascade, which is made up of BK, NO, and cGMP, AT2R counteracts the AT1R-mediated vasoconstrictor action of Ang II and promotes vasodilation activity. High Ang II causes the AT2R to further promote the repression of renin production and secretion, which indirectly offsets the negative effects caused by the AT1R [24, 33] . In the current investigation, these two AT2R-mediated inhibitory effects have been included in the traditional RAS model to evaluate how this feedback control mechanism controls the MAP.

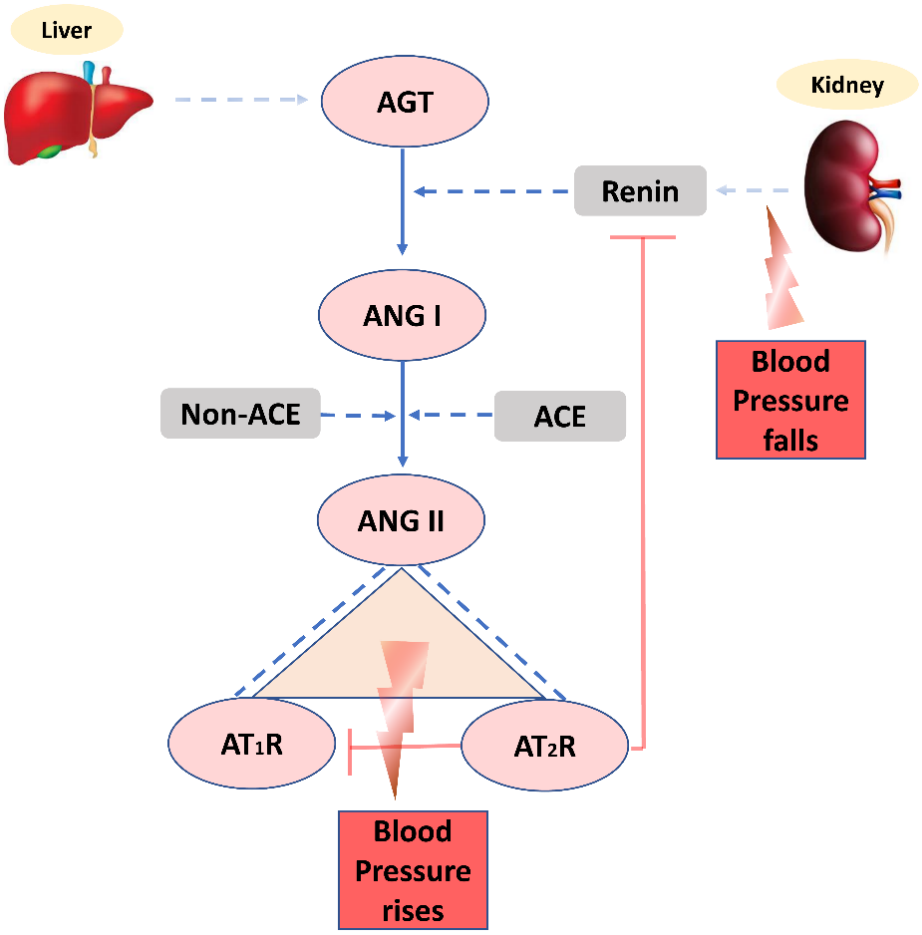


Figure 2.2 Block diagram of the renin-angiotensin (cRAS) system

Renin, an enzyme produced by the kidneys in reaction to low blood pressure, cleaves circulating angiotensinogen (AGT) in the blood secreted from the liver into angiotensin I (ANG I). Angiotensin-converting enzyme (ACE) and non-ACE both work together to transform this inactive peptide, ANG I, into the primary effector component, ANG II. The angiotensin type-1 receptor (AT1R) and the angiotensin type-2 receptor (AT2R) are two G-protein-coupled receptors that are bound by this ANG II. Vascular contractions that result in elevated blood pressure are caused by ANG II binding to the AT1R and activating cytoplasmic signalling pathways. While ANG II coupled to the AT2R blocks the AT1R's effects and reduces renin's ability to synthesise. The pink line denotes these negative feedback effects of AT2R.

2.2.1 Mathematical model of cRAS

The time-dependent concentration (mol/L) changes of all cRAS components are mathematically described by the traditional RAS model, which is expressed by a set of ordinary differential equations (ODEs). Six of the seven equations [34, 35] incorporate cRAS components; however, the final equation[36] describes fluctuation of mean-arterial pressure (MAP) based directly on Ang II concentrations. Each ODE represents a mass balance with a fixed volume. In relation to other cRAS components, the rates of creation, consumption, and degradation are used to explain the rates of changes in concentration. Production of angiotensinogen (AGT) follows zero-order with a rate constant of kAGT. First-order kinetics uses the half-lives, h_i, of the species *i* to define the rate of degradation. First-order kinetics are thought to govern the enzymatic and binding reactions, with enzyme *i's* rate constants c_i. Two additional inhibitory feedback effects mediated by AT2R have been incorporated into this conventional model: i) that proportionately changes with the concentrations of Ang II-bound AT2R, with the proportionality constant k_1 lowering the free availability of Ang II for any time t, and ii) that suppresses the renin concentration which is presumed to be jointly proportional to the both Ang II-bound AT1R and AT2R concentrations with the proportionality constant k_2. The final cRAS-ODE systems with the AT2R-induce feedback effects are as follows:

$$\frac{d[AGT]}{dt} = \text{kAGT} - \text{cRenin}[\text{AGT}] - \frac{ln2}{hAGT}[\text{AGT}] \quad 2.1$$

$$\frac{d[Renin]}{dt} = sRenin + \text{k}_f([\text{ANGII}]_0 - [\text{ANGII}])\left(1 - \frac{[ANGII]_0 - [ANGII]}{f}\right) - \frac{ln2}{hRenin}[\text{Renin}] - k_2\,[AT_1\text{R\_ANGII}][AT_2\text{R\_ANGII}] \quad 2.2$$

$$\frac{d[ANGI]}{dt} = cRenin[AGT] + kRenin([Renin] - [Renin]_0) - cAItoll[ANGI] - \frac{ln2}{hANGI}[ANGI] \quad 2.3$$

$$\frac{d[ANGII]}{dt} = cAItoll[ANGI] - \left(CAT1 + CAT2 + \frac{ln2}{hANGII}\right)[ANGII] \quad 2.4$$

$$\frac{d[AT_1R\_ANGII]}{dt} = \text{CAT1}[ANGII] - \frac{ln2}{hAT1\text{R\_ANGII}}[AT_1R\_ANGII] - k_1[AT_2R\_ANGII] \quad 2.5$$

$$\frac{d[AT_2R\_ANGII]}{dt} = CAT2[ANGII] - \frac{ln2}{hAT_2R\_ANGII}[AT_2R\_ANGII] \quad 2.6$$

$$\frac{dMAP}{dt} = kMAP[ANGII] - gamma[MAP] \quad 2.7$$

2.2.2 Canonical model for circulating-RAS

1. $\frac{d[AGT]}{dt}$=KAGT- cRenin [AGT] - $\frac{ln2}{hAGT}$[AGT],

where KAGT denotes the steady rate of AGT production. cRenin is the rate parameter for the conversion of AGT to ANGI that is catalysed by renin. AGT's half-life degradation is known as hAGT.

2.

$$\frac{d[Renin]}{dt} = sRenin + \text{kf}([\text{ANGII}]0 - [\text{ANGII}])\left(1 - \frac{[ANGII]0 - [ANGII]}{f}\right) - \frac{ln2}{hRenin}[\text{Renin}] - k2\,[AT1\text{R\_ANGII}][AT2\text{R\_ANGII}]$$

where, sRenin is the constant source of renin in absence of feedback. The second term

is the influence of ANGII negative feedback on renin production. The last term is the influence of negative feedback by the combined effect of AT1R and AT2R on renin production. hRenin is the half-life degradation of renin.

3.

$$\frac{d[ANGI]}{dt} = cRenin[AGT] + kRenin([Renin] - [Renin]_0) - cAItoll[ANGI] - \frac{ln2}{hANGI}[ANGI]$$

where the first term denotes the contribution of renin-catalyzed ANGI synthesis from AGT. The second term depicts the switch from AGT to ANGI synthesis as a result of ANGII's feedback on renin, which has a rate constant of kRenin. Third term is the cAItoII is the catalysed conversion of ANGI to ANGII. hANGI is the half-life degradation of ANGI.

4. $\frac{d[ANGII]}{dt} = cAItoll[ANGI] - \left(CAT1 + CAT2 + \frac{ln2}{hANGII}\right)[ANGII]$

The first term is the ANGII production. ANGII is consumed in the second term after binding with AT1R and AT2R. The half-life degradation of ANGII is hANGII.

5. $\frac{d[AT1R\_ANGII]}{dt} = CAT1[ANGII] - \frac{ln2}{hAT1R\_ANGII}[AT1R\_ANGII] - K[AT2R\_ANGII]$

The first term is the production of $AT1R\_ANGII$ from ANGII. h$AT1R_ANGII$is the half-life degradation of $AT1R\_ANGII$. The last term is the influence of negative feedback by the AT2R on $AT1R\_ANGII$ production.

6. $\frac{d[AT2R\_ANGII]}{dt} = CAT2[ANGII] - \frac{ln2}{hAT2R\_ANGII}[AT2R\_ANGII]$

The first is the production of $AT2R\_ANGII$ from ANGII. h$AT2R_ANGII$ is the half- life degradation of $AT2R\_ANGII$.

7. $\frac{dMAP}{dt} = kMAP[ANGII] - gamma[MAP]$

The Mean Arterial Pressure, or MAP. The PAM constant is kMAP. The PAM decay rate is gamma. *Peptidylglycine-α-amidating monooxygenase* (PAM) may play a role in the secretion of atrial natriuretic peptide (ANP), which is a hormone involved in the maintenance of blood pressure (BP).

The list of parameters, units, their definitions, and numerical values with sources is below:

Table 2.1 List of cRAS model parameters, together with definitions and references to standard values for each

Model Parameters	**Unit**	**Parameters Definition**	**Numerical Values**	**References**
KAGT	mol/L/s	Constant production rate of AGT	6.3×10^{-7}	[34]
hAGT	s	The half-life degradation of AGT	3.6×10^{4}	[34]
Renin0	mol/L	Initial concentration of Renin	2.06×10^{-13}	[37]
sRenin	mol/L/s	The constant source of renin in the absence of feedback	9.519×10^{-16}	[37]
hRenin	s	The half-life degradation of Renin	15	[37]
kRenin	s^{-1}	Feedback rate constant	17.89	[37]
cRenin	s^{-1}	Rate constant for the production of ANGI from AGT	1.7×10^{-14}	[34]
CAItoll	s^{-1}	Reaction rate constant for the parallel pathways catalyzing the conversion of ANG I to ANG II	6.7×10^{-3}	[34]
CAT1	s^{-1}	The Glucose dependent rate parameter for binding ANGII to AT1R receptor	1.4×10^{-2}	[34]
CAT2	s^{-1}	The Glucose dependent rate parameter for binding ANGII to AT2R receptor	1.2×10^{-2}	[34]
hANGI	s	The half-life degradation of ANGI	0.62	[38]
hANGII	s	The half-life degradation of ANGII	18	[38]
hAT1R_ANGII	s	The half-life degradation of hAT1R_ANGII	1.5	[38]
hAT2R_ANGII	s	The half-life degradation of Renin hAT2R_ANGII	1.5	[38]
kMAP	mmHG M^{-1} s^{-1}	PAM constant	0.05×10^{10}	[39]
Gamma	s^{-1}	PAM decay rate	0.15×10^{-3}	[39]
Kf	s^{-1}	Feedback parameter	1.36×10^{-8}	[34]
F	mol/L	Feedback parameter	5.04×10^{-7}	[34]
	mol/L	Initial concentration of ANGII	21×10^{-9}	[34]
k_1	s^{-1}	Negative feedback control parameter	Estimated	In this study
k_2	s^{-1}	Negative feedback control parameter	Estimated	In this study

2.3 Numerical solutions of non- linear ODEs

Ordinary differential equations (ODEs) are often solved numerically using the fourth-order Runge-Kutta technique. It falls under the category of an explicit technique, which indicates that only data from earlier time steps are used to determine the answer at a particular time step. For,

$$\frac{dy}{dx} = f(x, y), y(0) = y_0 \tag{2.8}$$

Runge Kutta 4th order method is given by

$$y_{i+1} = y_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)h, \tag{2.9}$$

where, $k_1 = f(x_i, y_i)$ (2.10)

$$k_2 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_1 h\right) \tag{2.11}$$

$$k_3 = f\left(x_i + \frac{1}{2}h, y_i + \frac{1}{2}k_2 h\right) \tag{2.12}$$

$$k_4 = f(x_i + h, y_i + k_3 h) \tag{2.13}$$

The function *ode45* is used in MATLAB which serves as the default solver for ODEs. For effective computing, this function uses a Runge-Kutta method with a configurable time step. *ode45* is made to address the following general problem:

$$\frac{dx}{dt} = f(t, x), \ \mathbf{x}(t_0) = x_0, \tag{2.14}$$

where t is the independent variable, x is a vector of dependent variables to be determined, and f (t, x) is a function of t and x. When the vector of functions on the right-hand side of Eq. (2.14), f (t, x), is set and the initial conditions, $x = x_0$ at time t_0, are given, the mathematical model is identified.

2.4 Non-linear least square method

The non-linear least squares method is a mathematical optimization technique used to find the parameters of a non-linear function that best fit a given set of data. Given a set of data points (x_i, y_i) where i=1,2,...,n, and we want to find the parameters a, b, c,... that best fit the following non-linear model:

$$y = f(x; a, b, c,...), \qquad 2.15$$

where f is a non-linear function with the parameters a, b, c, …

The goal of the non-linear least squares method is to find the values of the parameters a, b, c,... that minimize the sum of the squared differences between the observed data points and the corresponding values predicted by the model. This can be formulated as an optimization problem as follows:

$$\text{minimize } S(a, b, c, ...) = \sum[y_i - f(x_i; a, b, c, \ldots)]^2, \qquad 2.16$$

where the summation is taken over all the data points.

The optimization problem is created using the default properties when *prob = optimproblem* is used. The additional parameters given by one or more Name, Value pair inputs are used by the formula *prob = optimproblem(Name,Value).* Use *prob = optimproblem('ObjectiveSense','maximize')* to describe a maximization problem rather than a minimization problem, for instance.

2.5 *lsqnonlin* function

It is a particular variety of nonlinear least-squares solver that deals with nonlinear least-squares curve fitting issues of the following form.

$$\min_x \|f(x)\|^2_2 = \min_x (f1(x)^2 + f2(x)^2 + ... + fn(x)^2) \qquad 2.17$$

with *lb* and *ub* optional lower and upper boundaries on the *x*-components.

x, *lb*, and *ub* may be matrices or vectors.

lsqnonlin needs the user-defined function to calculate the vector-valued function

f(x)= $\begin{bmatrix} f1(x) \\ f2(x) \\ . \\ . \\ fn(x) \end{bmatrix}$ rather than computing the value of $\|f(x)\|^2_2$(the sum of squares).

x= lsqnonlin (fun,x0) begins at *x0* and seeks the least of the sum of the squares of the *fun* functions. Instead of returning the sum of the values' squares, the function *fun* should return a vector (or array) of values. The optimization options listed in options lead to the minimization of *x = lsqnonlin(fun,x0,lb,ub,options).* To set these options, use *optimoptions*. If there are no boundaries, pass empty matrices for *lb* and *ub.*

2.6 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is a very effective method for figuring out whether two samples are significantly different from one another. It is typically used to verify the consistency of random numbers. Any random number generator should have uniformity, and the Kolmogorov-Smirnov test can be used to check for it. The

Kolmogorov-Smirnov statistic measures the difference between two samples' empirical distribution functions or between their empirical distribution functions and the cumulative distribution function of the reference distribution. This statistic's null distribution is calculated under the assumption that the sample was taken from the reference distribution (in the event of a single sample) or that all samples were taken from the same distribution (in the two-sample case).

For n independently dispersed and identically distributed (i.i.d.) ordered observations X_i, the empirical distribution function F_n is defined as

$$\mathbf{F_n(x)}=\frac{number\ of\ (elements\ in\ sample \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(-\infty,x]}\boldsymbol{X_i}, \qquad 2.18$$

where $1_{(-\infty,x]}X_i$ is the indicator function equal to 1 if $X_i \leq$ x and equal to o otherwise.

With respect to a specific cumulative distribution function F(x), the Kolmogorov-Smirnov statistic is

$$\mathbf{D_n}=\boldsymbol{sup_x}|\boldsymbol{Fn(x)} - \boldsymbol{F(x)}|, \qquad 2.19$$

where sup_x is the set of distances' supremum. The statistic, intuitively, selects the biggest absolute difference between the two distribution functions over all possible values of x.

2.7 Local sensitivity analysis

A technique called local sensitivity analysis is used to evaluate how changes in input variables would affect a model's or system's output. It entails examining how sensitive the model output is to little changes in the input variables close to a certain

point. A model's behavior under various circumstances may be understood and the most important input variables can be found by using local sensitivity analysis. Additionally, it can help with model validation and verification. The fact that local sensitivity analysis only offers information on the sensitivity of the model output to minor perturbations around a particular place is one of its limitations. It does not offer details on the model's overall behavior or the results of more significant perturbations. To completely comprehend the behavior of a model or system, it is crucial to integrate local sensitivity analysis with other methodologies, such as global sensitivity analysis. In conclusion, local sensitivity analysis is a beneficial approach for examining the influence of changes in input variables on the output of a model or system. It gives a gauge of each input variable's relative weight and can help with model validation and verification. To completely comprehend a model or system's behavior, it is crucial to combine local sensitivity analysis with other methods.

2.8 Synthetic data generation and in-silico Experiment

In order to investigate the impact of AT2R-mediated feedback regulation under various cardiovascular situations, the traditional RAS model is linked with changes in MAP over time. The formula MAP=DP+1/3(SP-DP) is frequently used to calculate the MAP, which stands for mean arterial pressure for a single cardiac cycle; DP stands for diastolic pressure, and SP for systolic pressure.

Three separate cardiovascular physiologic conditions have been taken into consideration, each of which is indicated by a different range of MAP variations: the normal situation (MAP 70–100 mmHg), hypertension (MAP 101–190 mmHg), and hypotension (MAP 40–69 mmHg). The *ode45* function in MATLAB version R2021a

was used to use the fourth-order Runge-Kutta method to numerically solve the ODE model. The model has been run for 10^6 seconds (or about 12 days), with consistent time increments of 0.01 seconds, giving it plenty of time to reach a feasible steady state.

The steady-state numerical solutions of the model's MAP variable are fitted against the randomly chosen MAP values in order to estimate the unknown feedback control parameters k_1 and k_2 and explore AT2R-mediated feedback effects. 100 randomly selected points are taken from the defined known ranges of the MAP fluctuation for each circumstance (normal, hypertension, and hypotension), and these points are then fitted with the model MAP steady-state solutions using the non-linear least square approach. The sum of squared differences near to zero indicates the appropriate fits. We have minimized the sum of squared differences between the randomly selected MAP values from the prescribed ranges and the model solution of MAP steady states in order to obtain the most accurate estimates of the parameters k_1 and k_2. The set objective function of the sum of squared differences was used with the MATLAB's *optimproblem* function to define and perform this optimization problem. Using the MATLAB *lsqnonlin* function, which enables the local minimum with an estimation of the sum of squared differences, this problem has been solved for the parameters k_1 and k_2 (**Figure 2.3**).

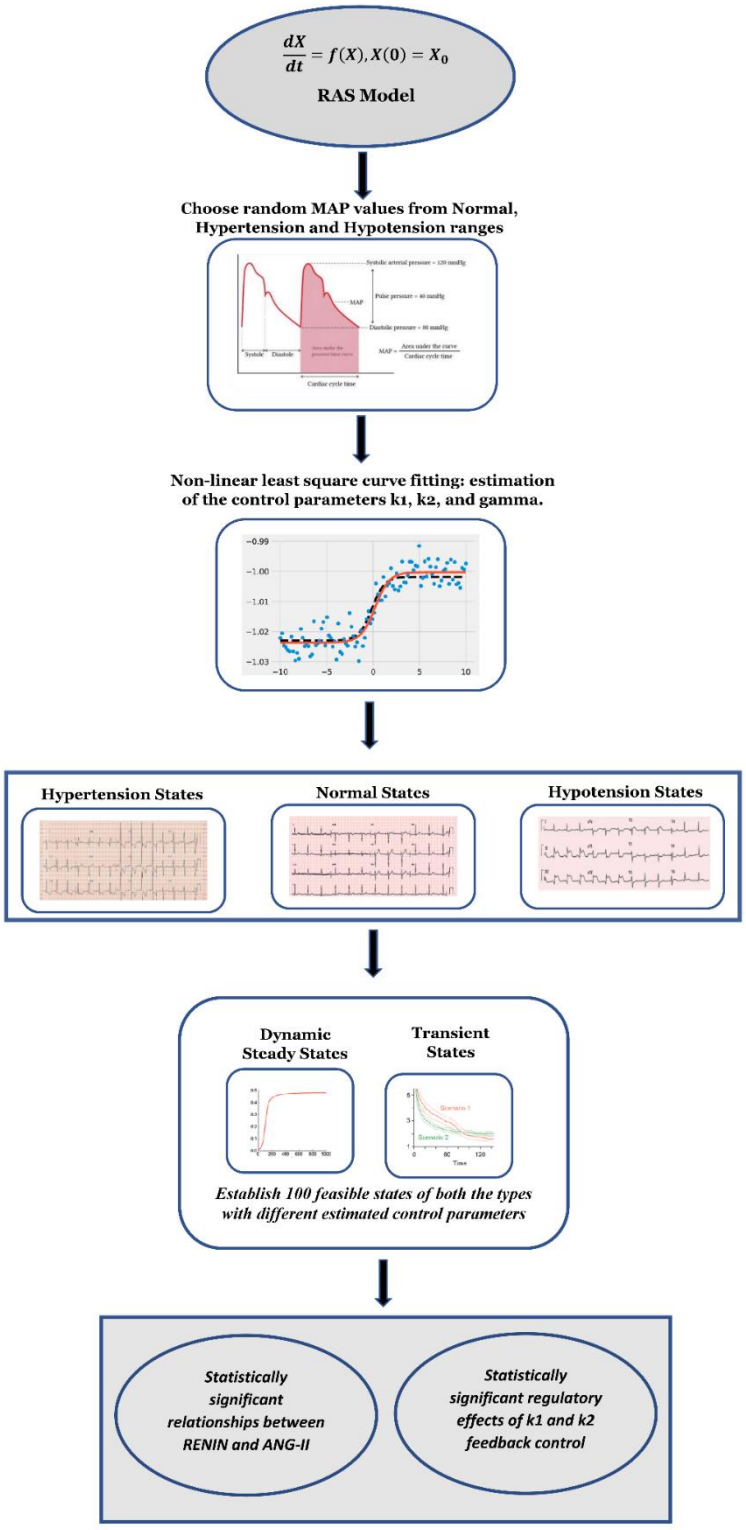


Figure 2.3 In-silico experiment flow chart

i) three disparate ranges of MAP variations are considered to defined three distinct cardiovascular physiologic conditions, 70-100 mmHg (normal), 101-190 mmHg (hypertension),

and 40-69 mmHg (hypotension); (ii) Randomly picked MAP values from these ranges are fitted against the model steady-state solutions of MAP variables and minimized the sum of squared differences; (iii) Three states are established: normal, hypertension and hypotension states with the estimated control parameters values of k_1 and k_2; (iv) 100 dynamic steady states for each condition (normal, hypertension, and hypotension) are established to examine their relationship and feedback effects and the mean transient trajectories are also evaluated, (v) Statistical significance of observed differences in the model simulated samples are calculated using Kolmogorov-Smirnov (KS) test and Q-Q plot.

For the three scenarios—normal, hypertension, and hypotension—three sets of parameter values are estimated. With the allowable sum of squared differences 0.01, we selected 100 estimated parameter values of k_1 and k_2 in each set to compare significant differences between these three cardiovascular states. With these parameter values, the model has been run once more, and steady-states have been established. According to the parameters' belongingness, this simulation created 100 steady states for each case. We have calculated average transient states and their standard deviations in addition to steady-states.

The Kolmogorov-Smirnov test has been used to assess pairwise differences between the normal, hypertensive, and hypotensive states, and p-values and D-statistics are derived from the model-simulated samples. Additionally, Q-Q plots are created to visually depict these variations, indicating which distribution is more skewed or which has heavier tails.

For the control parameters k_1 and k_2, a local sensitivity analysis was performed to assess the reliability of the feedback control. From the model-simulated samples, we have first evaluated the range of variations of k_1 and k_2 under normal, hypertension, and hypotension situations. Each parameter's lower and upper bounds

are calculated with a 95% confidence level using its mean and standard deviation. The steady-state concentrations of renin and Ang II are recalculated after the mean values of k_1 and k_2 are perturbed within 0.01-neighborhood. The steady states of renin and Ang II were examined in terms of percent changes. In order to quantify how much the degree of controls exerted by k_1 and k_2 varied depending on the condition, these results were compared among the normal, hypertension, and hypotension states. We have normalized the model solutions X using the formula,

$$X \to \frac{X - X_{min}}{X_{max} - X_{min}}$$

to make it into a unit less amount lying between 0 and 1 in order to compare the steady-state concentrations of renin and Ang II displaying very low numerical quantity (i.e., multiple of 10^{-7}) and to perform statistical testing.

2.9 Results

a) At steady states, differential AT2R-mediated feedback controls led to the emergence of a linear relationship.

To assess transient and steady-states across various cardiovascular physiologic situations (such as normal, hypertension, and hypotension) defined by differential ranges of MAP change, the model solutions of Ang II and renin are tracked through time. For each particular scenario, 100 model-simulated random samples were taken, each with a unique set of AT2R-mediated control parameters. It has exhibited considerable fluctuation in feedback control parameters in hypertension and hypotension situations compared to its normal states while retaining similar steady-

states of Ang II and renin (approximately $1.52 \times 10^{-7} \pm 3.8 \times 10^{-10}$, and $9.11 \times 10^{-8} \pm 2.2 \times 10^{-10}$ mol/L, respectively).

Renin achieves the steady state more quickly than Ang II. There are notable variances in the transient dynamics even though both concentrations monotonically rise and plateau. When compared to hypertension and hypotension, it separates transient states with considerably different average values over time (**Figure 2.4**). With normalized concentrations, the Q-Q plot (**Figure 2.5**) and the Kolmogorov-Smirnov test further illustrate these disparities (**Table 2.2**). At steady-states, the cRAS system with AT2R-mediated feedback controls upholds a linear relationship between the concentrations of renin, Ang-I, and II (**Figure 2.6**):

$$[\text{Renin}]^* = a_1[\text{ANG II}]^* + a_2[\text{ANG I}]^* + a_3 \qquad 2.20$$

where, $a_1 = \frac{\left(CAT1 + CAT2 + \frac{ln2}{hANGII}\right)}{k\ Renin}$

$$a_2 = \frac{ln2}{kRenin\ .\ hANGI}$$

$$a_3 = [\text{Renin}]_0 - \frac{cRenin[AGT]}{kRenin}$$

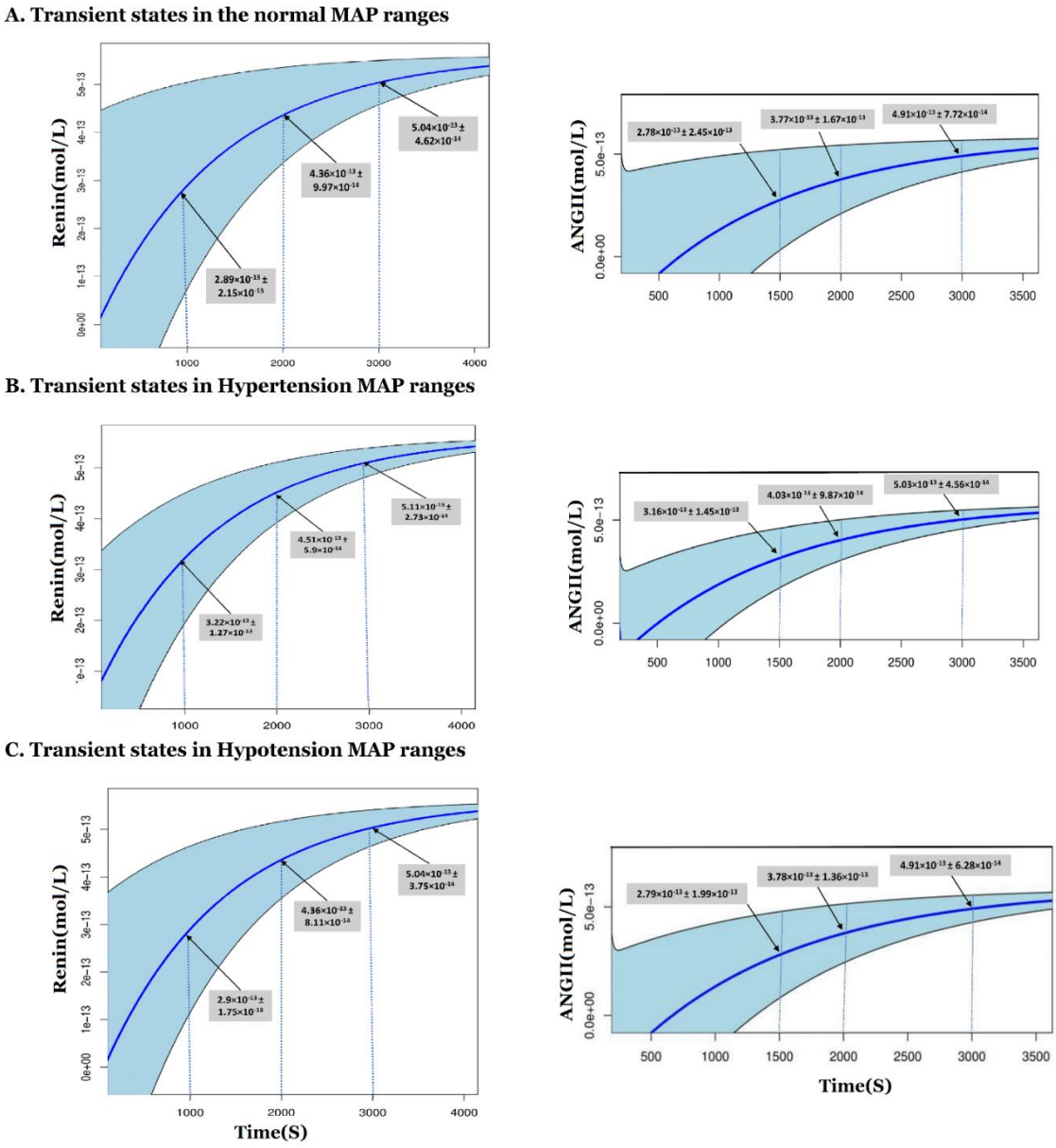


Figure 2.4 Transient dynamic of [Renin and [ANGII]

Transient dynamics of [Renin] and [ANGII]-mean trajectory (blue line) with standard deviations highlighted in cyan hue, (A) when the MAP varies in the normal range 70-100 mmHg, (B) the MAP varies in the range 101-190 mmHg (hypertension), and (C), the MAP varies in the range 40-69 mmHg (hypotension).

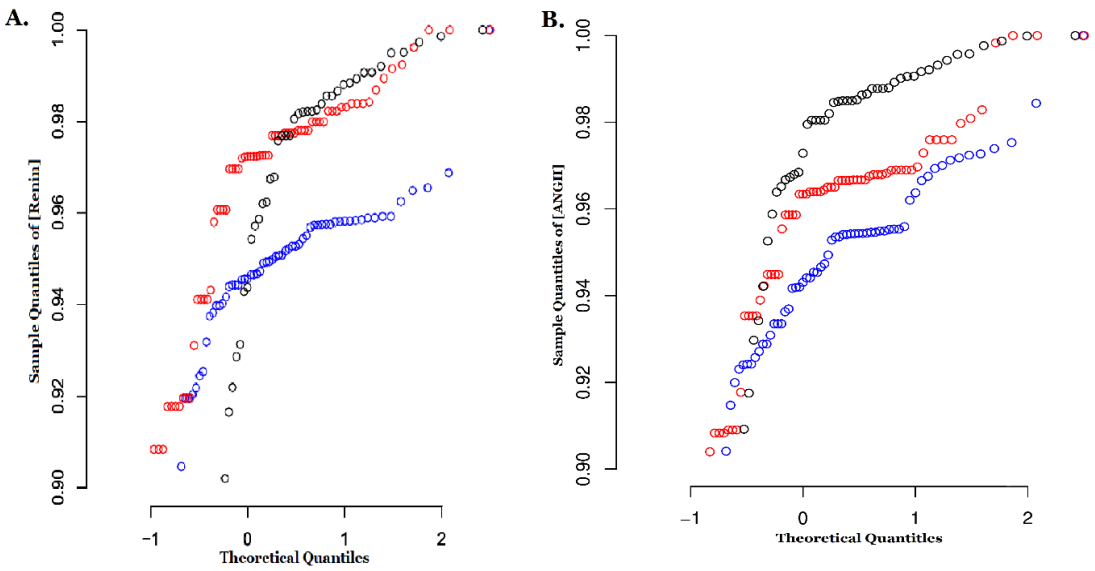


Figure 2.5 Q-Q plots showing deviations of [Renin] and [ANG II]

The normal condition is shown by a red ring, hypertension and hypotension are denoted by a blue ring, and the deviations of [Renin] and [ANG II] states from the normal state are shown by Q-Q plots, respectively.

Table 2.2 Kolmogorov-Smirnov (KS) test results of significant differences among the three MAP condition

KS-test	Dynamic States	Normal				Hypertension				Hypotension			
		p-value		D-statistic		p-value		D-statistic		p-value		D-statistic	
		[Renin]	[ANG II]	[Renin]	[ANG II]	[Renin]	[ANG II]	[Renin]	[ANG II]	[Renin]	[ANG II]	[Renin]	[ANG II]
Normal	**Steady-state**	**1**	**1**	**0**	**0**	**1.632×10^{-13}**	**2.2×10^{-16}**	**0.5650**	**0.6280**	**1.341×10^{-3}**	**1.216×10^{-7}**	**0.3080**	**0.4640**
	Transient state	**1**	**1**	**0**	**0**	**2.2×10^{-16}**	**2.2×10^{-16}**	**0.017696**	**0.022696**	**0.9987**	**7.14×10^{-7}**	**0.000537**	**0.003853**
Hypertension	**Steady-state**	**1.632×10^{-13}**	**2.2×10^{-16}**	**0.5650**	**0.6280**	**1**	**1**	**0**	**0**	**5.995×10^{-15}**	**2.2×10^{-16}**	**0.6600**	**0.7340**
	Transient state	**2.2×10^{-16}**	**2.2×10^{-16}**	**0.017696**	**0.022696**	**1**	**1**	**0**	**0**	**2.2×10^{-16}**	**2.2×10^{-16}**	**0.017246**	**0.022177**
Hypotension	**Steady-state**	**1.341×10^{-3}**	**1.216×10^{-7}**	**0.3080**	**0.4640**	**5.995×10^{-15}**	**2.2×10^{-16}**	**0.6600**	**0.7340**	**1**	**1**	**0**	**0**
	Transient state	**0.9987**	**7.14×10^{-7}**	**0.000537**	**0.003853**	**2.2×10^{-16}**	**2.2×10^{-16}**	**0.017246**	**0.022177**	**1**	**1**	**0**	**0**

b) The MAP variation is more significantly impacted by AT2R-mediated reduction of renin activity.

The findings demonstrate that there are considerable differences in AT2R-mediated feedback control across the normal, hypertensive, and hypotensive states (**Table 2.3**). It should be emphasised that different control parameter sets can produce similar cRAS steady states for each of the three MAP conditions that are being taken into consideration. It follows that this points to the existence of several AT2R-mediated regulation in both normal and hypertensive and hypotensive situations. The present work calculated these two separate sets of control parameters, k_1 and k_2, and found significant variations between them at their mean values and 95% confidence intervals. In particular, k_1 levels remained close to 0.008 ±0.005 under normal, hypertensive, and hypotensive circumstances. Therefore, it suggests that AT2R has consistently negative effects on AT1R across all MAP variabilities. The duration of the k_2 variation increases by ~4.6 and 2 times, respectively, in hypotension and hypertension comparing to the normal MAP conditions, with a typical range of (2.15, 2.67) in the normal MAP fluctuation. In contrast, k_2 has drastically different means and possible ranges of variations. This data shows that strict AT2R-mediated control is loosened in hypertension and hypotension, and substantially higher renin activity suppression is required to maintain normal MAP.

Table 2.3 Negative feedback control parameters in the RAS model in different map conditions

	Hypertension	**Hypotension**	**Normal**
k_1 **(95%CI)**	0.0067 - 0.0092	0.0090 - 0.0104	0.0068 - 0.0091
k_2 **(95%CI)**	1.5670 - 2.4572	2.7642 – 4.9295	2.1503 - 2.6666
Mean k_1	0.0080 ± 0.0055	0.0097 ± 0.0027	0.0080 ± 0.0051
Mean k_2	2.0121 ± 2.0057	3.8468 ± 4.0590	2.4084 ± 1.1706

c) The interactions of the cRAS components led to the emergence of robust control of MAP.

The receptors AT1R and AT2R mediate cRAS responses to hypertension and hypotension as well as its feedback effects on MAP variation, inducing cascaded actions that alter important cRAS regulatory components[9]. The findings demonstrated that varied ranges of MAP variation influence AT2R mediated negative feedback effects on AT1R and on the renin production and secretion. Whether these modifications are extremely sensitive or exhibit a certain level of robustness over an extended length of time has been confirmed. The feedback parameters k_1 and k_2 have undergone local sensitivity analysis in order to answer this query. The findings of this investigation demonstrate that small perturbation in both k_1 and k_2 in normal, hypertensive, and hypotensive MAP variations cannot appreciably alter concentrations of renin and ang II (**Table 2.4**). The unaltered linear connections between renin and ang concentrations that are maintained despite MAP fluctuations also reflect these sensitivity (**Figure 2.6**).

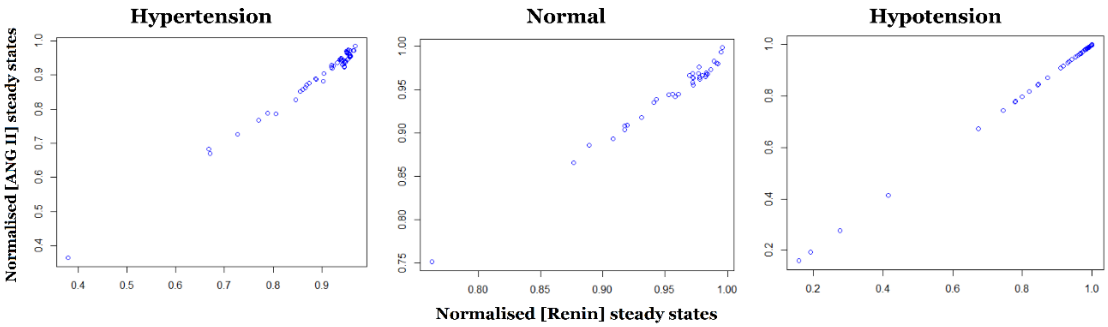


Figure 2.6 A linear steady-state relationship between [Renin and [ANGII]

Renin and Ang II have a linear connection in steady state. It should be noted that the AT2R-mediated feedback mechanism that enables cRAS to maintain this linear relationship differs dramatically under normal, hypertensive, and hypotensive situations.

Table 2.4 Local sensitivity analysis of AT2R-mediated feedback control parameters

Normal				**Hypertension**				**Hypotension**			
AT2R-mediated Feedback control parameters		**[Renin]***	**[ANG II]***	**AT2R-mediated Feedback control parameters**		**[Renin]***	**[ANG II]***	**AT2R-mediated Feedback control parameters**		**[Renin]***	**[ANG II]***
k_1=0.0080	k_2=2.408	Fixed	Fixed	*k1*=0.0080	*k2*=2.0121	Fixed	Fixed	*k1*=0.0097	*k2*=3.8468	Fixed	Fixed
k1=0.008+ 0.01	*k2*=2.408+ 0.01	0.0020	0.0020	*k1*=0.0080+ 0.01	*k2*=2.0121+ 0.01	0.00157	0.00157	*k1*=0.0097+ 0.01	*k2*=3.8468+ 0.01	0.003501	0.003501
k1=0.008 - 0.01	*k2*=2.408-0.01	0.0020	0.0020	*k1*=0.0080- 0.01	*k2*=2.0121- 0.01	0.00155	0.00155	*k1*=0.0097- 0.01	*k2*=3.8468- 0.01	0.003479	0.003479

2.10 Summary

a) In this Chapter I have discussed the adverse feedback effects brought on by AT2R-Ang II binding reactions serve as reliable controls on mean-arterial variance. Depending on the degree of the deviation, this control's dysregulation might result in either hypertension or hypotension.

b) Further it has been demonstrated that activating AT2R-renin interactions that suppress renin activity significantly contribute to the maintenance of healthy MAP fluctuation, which further suggests that this interaction may be a possible therapeutic target for lowering blood pressure.

"Spatiotemporal data holds the key to unlocking new insights into the relationships between space, time, and events" …… John Krumm

Chapter 3

3. Spatiotemporal data modelling and analysis-I

Graphical View

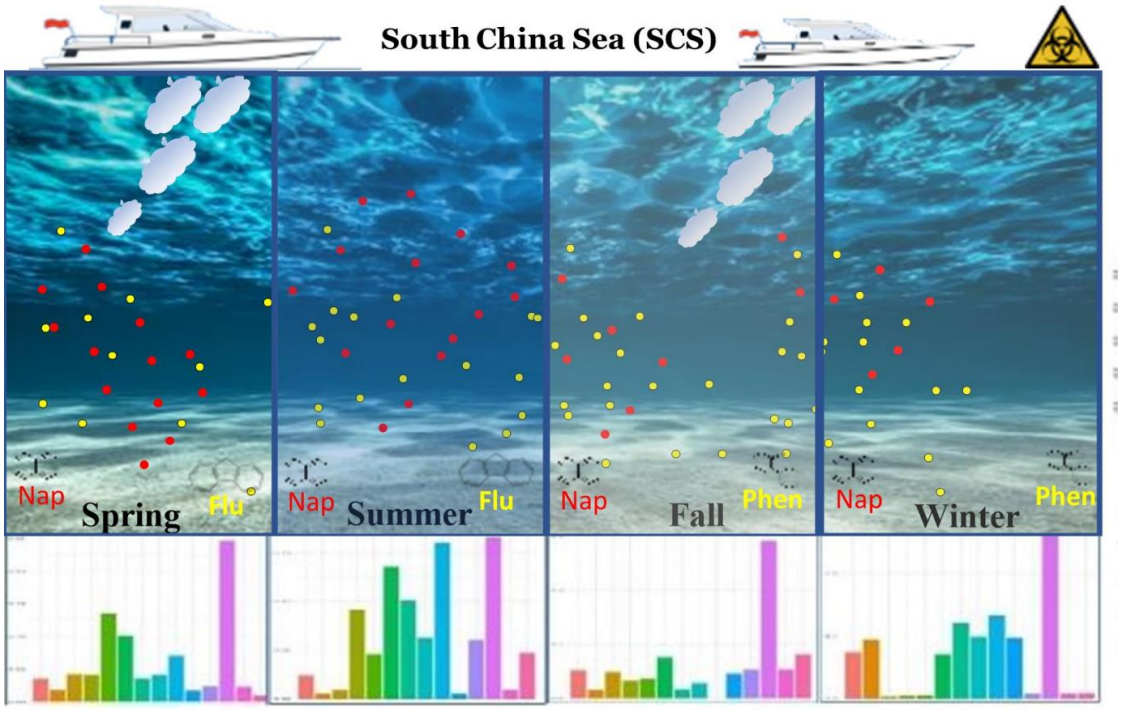


Figure 3.1 Graphical Abstract

Wang, C., **Thakuri, B.**, Roy, A. K., Mondal, N., Chakraborty, A. (2022). Phase partitioning effects on seasonal compositions and distributions of terrigenous polycyclic aromatic hydrocarbons along the South China Sea and East China Sea. *Science of The Total Environment, 828*, 154430.

3.1 Introduction

In the global biogeochemical cycles, a class of persistent organic pollutants known as polycyclic aromatic hydrocarbons (PAHs) is widely spread and is substantially to blame for the contamination of the environmental matrix[40-42]. Due to the strong mutagenic and carcinogenic properties of PAHs, which are a major hazard to human health and have negative effects on ecosystems, study interest has been focused on their occurrences and distributions for a number of decades[40, 43-46]. These effects are primarily caused by human activity, which results in pollution from incomplete combustion and pyrolysis of fossil fuels or wood (pyrogenic) or from the release of crude and refined petroleum (petrogenic), as well as from increased industrial activities that have an adverse effect on the environment[47-50]. When pervasive organic pollutants are released into the atmosphere, they are eventually transported by surface runoff and atmospheric transport processes and deposited in neighbouring marine habitats, where they bioaccumulate over time and reach high concentrations[40, 50]. Due to their hydrophobic characteristics, released PAHs are either immediately absorbed by suspended particulate matter or dissolved and dispersed through the sea surface water[40]. Numerous PAH species cannot be deposited as sediment due to strong ocean currents. Recent studies have revealed substantial levels of bioaccumulation by marine life, which offers a long-term serious risk due to eco-toxicity, a human mutagenic and carcinogenic effect, retention in food chains, and an impact on ocean biogeochemical cycles[40, 51, 52] . Understanding the phase partitioning, geographical and temporal variations in composition, and distribution of PAH species is crucial for better coping with PAH contamination in ocean systems. This study addressed this issue and shown that the two phases of

dissolved and particulate forms along the continental shelf borders of the northern South China Sea (SCS) and East China Sea differ significantly in terms of PAH species composition and distribution (ECS).

Human activity has a significant impact on the marginal seas that separate continents from oceans, which modifies the source-sink dynamics of organic pollutants by serving as both a “source” and a “sink” for releasing terrigenous substances into oceans[53, 54]. Because most PAHs are semi-volatile, they frequently redistribute in dissolved and particulate forms, which creates significant uncertainty in forecasting their environmental fate[55, 56]. Only thorough sampling strategies with sound methodologies capable of explaining such variances will be able to reduce these inherent uncertainties. The SCS containing an area of 3,500,000 km2, located at the confluence of the Indian and Pacific Oceans[57], and the ECS covering an area of 770,000 km2, located at the confluence of the Sea of Japan and SCS, have both been taken into consideration for this study’s analysis of the issue [58]. Rapid population and economic growth in the neighbouring developing nations is causing more surface runoff and organic pollution deposition in the SCS and ECS. Additionally, the release of PAH is influenced by oil spills from oil tankers and shipwrecks in the water[59]. PAH contaminations in SCS or ECS have been the subject of a few recent investigations, however phase-wise spatial and temporal fluctuation of PAHs is frequently disregarded. The goals of this study are to describe seasonal spatially and temporal fluctuations and to phase petition along the water depth, both of which are crucial for assessing the ecological danger of PAHs.

3.1.1 PAH species

The U.S. EPA classified 16 PAHs that are quantified in two phases, dissolved and particulate forms as priority pollutants out of the more than 200 species of PAH that have been found based on their carcinogenic, mutagenic, and teratogenic properties to better deal with PAH contamination. The following 16 PAH species are taken into consideration: Acenaphthylene (Ace), Acenaphthene (Acen), Fluorene (Flu), Anthracene (An), Benzo(a)anthracene (BaA), Chrysene (Chry), Benzo(b)fluoranthene (BbF), Benzo(k)fluoranthene (BkFA), Benzo(a)pyrene (BaP), Indeno (1,2,3-cd) pyrene (IP), Fluoranthene (Fluo), Phenanthrene (Phen), Dibenzo (a,h) anthracene (DBA), Benzo (g, hi) perylene (BgP) Pyrene (Py), Naphthalene (Nap).

The two groups of these PAHs are divided based on their molecular weight: a) PAHs with a low molecular weight and fewer than four aromatic rings (i.e., Nap, Ace, Acen, Flu, Phen, An), and b) PAHs with a high molecular weight and four or more rings (i.e., Fluo, Py, BaA, Chry, BbF, BkFA, BaP, IP, DBA, and BgP). With its six aromatic rings, indeno (1, 2, 3-cd) pyrene (IP) has the highest molecular weight, coming in at 276.3 g/mol, while naphthalene (Nap), which only has two rings, has the lowest molecular weight, coming in at 128.1 g/mol. Compared to low molecular weight PAHs, these high molecular weight PAHs often have a lower water solubility and partition more easily into organic materials. It is crucial to independently compute the spatiotemporal distributions for the dissolved and particulate phases because these variations in water solubility and molecular weights are reflected in the spatiotemporal distribution and changes in PAH compositions.

3.2 Methods and materials

3.2.1 Study area and sample collection

A field survey was conducted between 2009 and 2011 on the northern SCS and ECS continental shelf margins using the scientific research ship “Dongfanghong-2”. The current study ran four water sampling operations in the spring of 2011 (April–June), summer of 2009 (July–August), fall of 2010 (October–December), and winter of 2010. (December 2009 to January 2010). The sampling stations of the four journeys had slightly varying settings due to the restrictions of the cruise routes and weather, but overall they covered the entire northern SCS and the ECS. The investigation has set up a few “overlapped” stations and sections in the Yangtze River Estuary and the Pearl River Estuary over the course of four seasons to ensure the consistency of the study area and the comparability of the data. Using a submersible pump or a stainless-steel drum (10L) and a stainless-steel drum with a volume of 80L, 30–50L of surface seawater (with a depth of no more than 1m) was collected at selected stations, and the volume was recorded. The saltwater was then filtered via a particle phase filter membrane and into a 4L brown bottle using a peristaltic pump or glass-fibre filter (GF/F) system, which was then utilised to remove the membrane from the filtered water sample. A mixture of 50 ng of five deuterated PAHs was added to a portion of the filtered water sample (between 8 and 12 litres) to serve as a substitute for the real standards. With a flow rate of roughly 1 mL/s through the extraction column and under vacuum, the solid-phase extraction system was employed to absorb and fix PAHs and other dissolved organic compounds. Prior to water sample enrichment, the solid-phase extraction (SPE) column was activated with 10 mL of methanol and eluted with 10 mL of ultrapure water (resistivity of 18.25Mcm). The acquired

granular phase GF/F and dissolved phase SPE column samples were then sealed in a small bag after being wrapped in pre-fired aluminium foil sheets. It was frozen and kept at -4°C. Referred to were the thorough sample analysis techniques.

3.2.2 Data analysis

Seasonal data were gathered from various sites at various water depths (spring, summer, fall, and winter) (WD). There were 92 sites in the spring, with WD varying from 0 to 4795 metres, and 109, 75, and 88 sites for the summer, fall, and winter, respectively, with WD varying from0 to 4795 metres, 0 to 3848 metres and 0 to 4180 metres respectively (**Figure 3.2**). By examining 29 overlapping sites for the particulate PAHs and 31 dissolved overlapping sites for all seasons, the seasonal heterogeneities were ascertained. The temperature of the water varied significantly by season at each site (spring: 24.91°C±3.77; summer: 28.56°C±1.58; fall: 22.37°C±3.44; winter: 19.11°C±5.19). All the locations' and all the seasons' water salinity (%) stayed almost the same (spring: 33.39±1.09; summer: 31.93±2.72; fall: 33.25±1.55; winter: 33.49±1.32). All the locations also assessed suspended particulate matter (mg/L), which varied greatly from one site to the next and seasonally, with an average range of 9–15 mg/L. For the 16 PAH species in both the dissolved and particulate phases, the Mahalanobis distance (MD) matrix and standard Person correlation matrix, respectively, were used to calculate the correlation and distances among the individual PAH concentrations in the R Programming Language (version 4.1.1) (); (The R Core Team, 2021)[60, 61], respectively. Non-metric multidimensional scaling (NMDS) was used to find the hidden dimensions in the data. The R programming language’s vegan package was used for this investigation (Jari Oksanen et al., 2020). The current work used the venerable Principal Component

Analysis (PCA) in MATLAB (version R2021a) to ascertain the significant degree of seasonal variation of various PAH species concentrations[62].Before doing the PCA, it was discovered that all of the seasons' unique PAH concentrations were intermingled across all of the overlapping locations independently for the dissolved and particulate PAH.

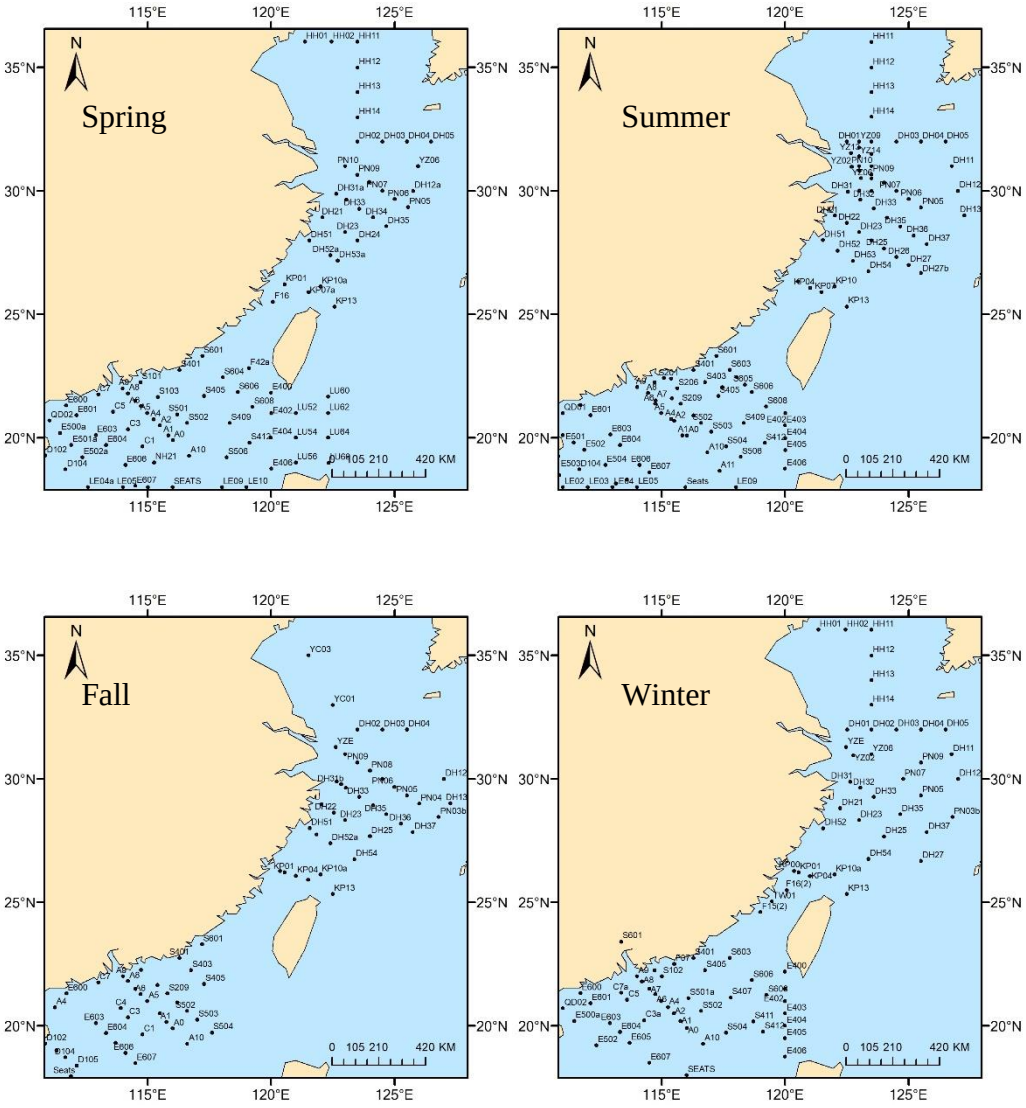


Figure 3.2 Sampling location along the South and East China Seas

Distribution map of the four seasons' sample locations around the South China Sea (SCS), which covers an area of 350 km^2. There are 31 and 29, respectively, total overlapping sites for dissolved and particulate PAHs.

3.2.2.1 Mahalanobis distance (MD)

The MD calculates the distance between a point and a distribution by counting the standard deviations between the point's location and the distribution's mean. As one gets away from the distribution's mean, the distance grows from zero. The MD i for the two random points $\vec{x}=(x_1,x_2,x_3,...,x_n)$ and $\vec{y}=(y_1,y_2,y_3,...,y_n)$ of the same distribution and the covariance matrix S, and x_i,y_i are the observation at i^{th} site, is defined by Eq. (3.1):

$$d(\vec{x},\vec{y})=\sqrt{(\vec{x}-\vec{y})s^{-1}(\vec{x}-\vec{y})} \qquad 3.1$$

3.2.2.2 Pearson's correlation matrix

If changes in one variable result in changes in the other, the two variables are said to be correlated. In general, correlation describes the statistical relationship between two variables. The Pearson correlation coefficient (r_{xy}) may be used to determine if two variables x and y have a linear relationship.

$$\mathbf{r_{xy}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\,(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \qquad 3.2$$

A correlation matrix is a table that shows the correlation coefficients for various variables. The correlation between all potential pairings of values in a table is shown in the matrix. It is an effective tool for finding and displaying trends in the provided data, as well as for summarising a huge dataset.

3.2.2.3 Non-metric multidimensional scaling (NMDS)

A non-metric multidimensional scaling (NMDS) technique si used to find the data's hidden dimensions. This ordination approach is unrestricted and makes no assumptions about how the data are distributed. There are no hidden axes of variation in NMDS; instead, a small number of axes are explicitly selected before the study and the data are fitted to those dimensions. Second, the majority of other ordination techniques are analytical, producing a single, distinctive answer to a collection of data. NMDS, on the other hand, is a numerical method that iteratively searches for a solution and terminates computation when a workable` solution has been identified, or after a certain number of attempts. The MDS technique is simple in concept but computationally challenging to implement. One first begins with a data matrix made up of p columns of variables and n rows of samples. Based on this, a n × n symmetrical matrix of all pairwise distances between samples is produced using a suitable distance measure, such as the Bray, Manhattan, and Euclidean distances. This distance matrix will be used for the MDS ordination. Next, the ordination's desired number of m dimensions is selected. The two ordinations would have to be carried out independently since a n-dimensional ordination is not comparable to the first n dimensions of an n+1-dimensional ordination.

A quantitative metric of ordination fit known as "stress", with a value of less than 0.15 representing goodness of fit, was used to determine the success of NMDS in dimensionality reduction[63] which can be calculated as,

$$Stress(1) = \sqrt{\sum_{h,i}(d_{hi} - \hat{d}_{hi})^2 / \sum_{h,i} d_{hi}^2} \quad 3.3$$

where $\hat{d}$ is the distance predicted by the regression and d_{hi} is the ordinated distance between samples h and i. Using the R programming language's vegan package, this analysis was carried out[64].

3.2.2.4 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a technique used for dimensionality reduction in multivariate data analysis. The main goal of PCA is to transform the original data into a new set of variables, called principal components (PCs), which capture most of the variation in the data with the least possible loss of information. PCA is based on the eigenvalue decomposition of the covariance matrix of the original data.

The PCA breaks down the complicated data into a smaller set of dimensions known as principal components (PCs), where each PC is independent of all other PCs[65, 66]. We may discover the data's hidden dimensions using this analysis, and PCA plots reveal probable clusters. Typically, the variability of the set of retained PCs, which is determined by the percentage of total variance that each PC accounts for, is used to assess the quality of the PCA which is given by Eq. (3.4),

$$\pi_j = \frac{\lambda_j}{tr(S)} \quad 3.4$$

where λ is the matrix's eigenvalue and $tr(S)$ represents the trace of the covariance matrix.

3.3 Results

3.3.1 Spatial distribution and PAH phase composition

Compared to low molecular weight PAHs, high molecular weight PAHs typically have a lower water solubility and partition more easily into organic materials. It is crucial to independently compute the spatiotemporal distributions for the dissolved and particulate phases because these variations in water solubility and molecular weights are reflected in the spatiotemporal distribution and changes in PAH compositions. Total PAHs computed as the sum of the averages of the 16 PAHs across the overlapping sites show significant seasonal variations. The summer months had the greatest overall PAH concentrations, which were 76.34 ng/L in the dissolved phase and 67.43 ng/L in the particulate phase. The dissolved phases show significant seasonal fluctuations in the spring (24.5 ng/L), summer (76.34 ng/L), fall (44.91 ng/L), and winter (51.1 ng/L). In contrast, summer has a much higher value of 67.43 compared to spring, fall, and winter's relatively low and equivalent total PAHs concentrations. Indicating regional homogeneity, which often characterises the marginal seas trapped between the continents, the data on spatial distributions among the 31 overlapping sites indicated little variations between the sites for both phases (**Figure 3.3**). In both the dissolved and particulate phases, it is noteworthy that only a few numbers of PAH species spatially dominate throughout all the locations, while others retain comparatively lower and almost equal concentrations. All of the SCE and ECS sites exhibit strong seasonal variation, and the between-phase changes continue to be very noticeable. Flu maintains at least a two-fold larger concentration in the dissolved phase than the other species during the spring and summer seasons. Flu showed its greatest mean concentration during the summer, which ranged from

34.48 ng/L to a wide 9.133-263.06 ng/L range. In contrast to the summer, which revealed a very broad range of variability in all the PAH species, other species showed a substantially reduced range of variations. Naphthalene (Nap), in contrast, continues to be common in the particulate phase despite having substantially lower concentrations than all other PAH species. Nap in particular has the greatest mean summer concentration of 54.78 ng/L, which is 10 times greater than its concentration in all other seasons. During summer, there is a very wide range of variance over all the overlapping sites, ranging from 4.05 to 324.2 ng/L. In contrast to the Nap, the other PAH compounds noticeably retain much lower particle concentrations throughout the year. While searching for compositional variations between the two phases and the four seasons, we discovered whole different compositional patterns. Flu's concentrations in the dissolved phase were at their maximum throughout the spring and summer, but during the fall and winter, Flu overtook Phen as the second most prevalent substance. The four seasons' PAH species compositions with concentrations greater than 1 ng/L are listed in the following order: Winter: Phen>Flu>Acen> Ace> An>Fluo>Py; Spring: Flu>Phen>Py>Acen>Fluo; summer: Flu>Phen> Ace>Acen>An>Fluo; fall: Phen>Flu>Acen>An>Py>Ace. Comparatively less PAH compounds had concentrations more than 1 ng/L in Nap particulate concentration, which remained the highest during all four seasons. Between PAH species, the compositional order has significantly varied. The only PAH species with quantities more than 1 ng/L in the spring, fall, and winter is nap. Several particulate PAH species retain concentrations greater than 1 ng/L over the summer, in the following order: Nap>Py>Phen>Fluo>Flu.

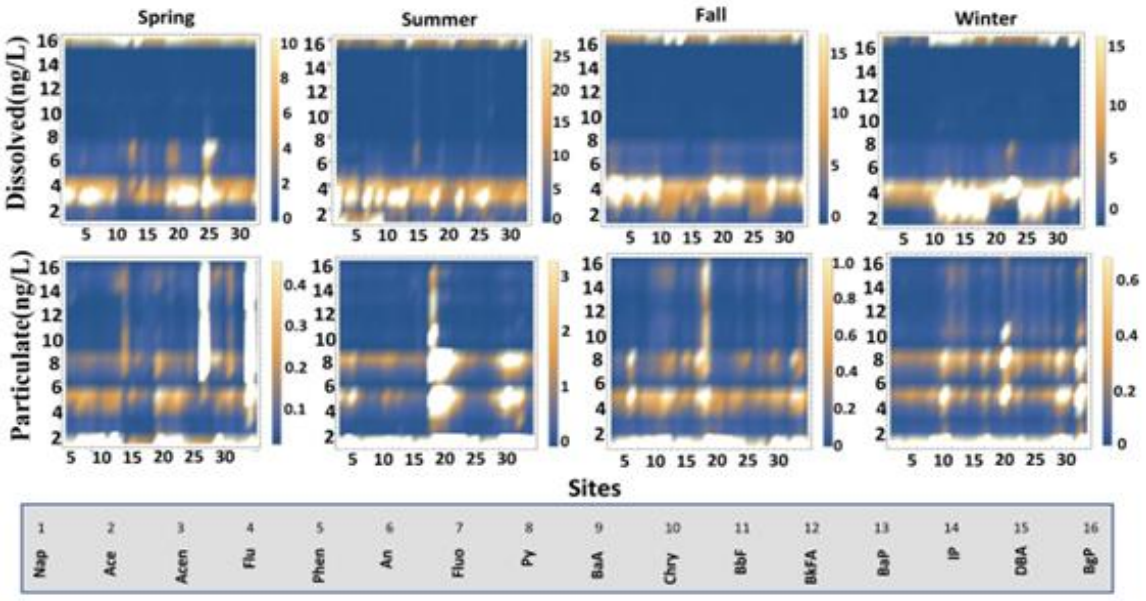


Figure 3.3 Spatial distribution of 16 PAH identified by U.S. EPA

It displayed the spatiotemporal fluctuation at 31 locations along the South China Sea (SCS), with varying concentrations throughout each of the four seasons: spring, summer, fall and winter, and two distinct measured types of concentrations (ng/l): dissolved and particulate.

3.3.2 Correlations and seasonal heterogeneity of PAH

We generated the pairwise Pearson correlation matrix and the MD matrix across all the overlapping locations across the four seasons in order to comprehend the spatiotemporal fluctuation of the 16 PAH species. The seasonal averages of PAH species were shown to have extremely high degrees of correlation and high MDs, which essentially explains the spatiotemporal patterns of persistently high concentration of a small number of PAH species. Notably, the findings revealed divergent association patterns between the amounts of dissolved and particulate PAH (**Figure 3.4 A)**. In general, there is a substantially stronger correlation between particulate PAH concentrations relative to dissolved PAHs. Strong correlations exist between consistently high particle concentrations of the Nap and Ace, Acen, Flu,

Phen, An, Fluo, Py, Chry, and BkFA (correlation coefficient >0.5, p-value < 2.2×10^{-16} The dissolved concentrations of Flu and Phen, which remained at greater levels across all four seasons, however, showed only a very weak connection with the concentrations of the other dissolved PAH species ($p<0.05$). As a result, it suggests that several oceanic factors (such as current, salinity, and water depth) may be to blame for the variance in concentration of dissolved and particulate PAHs. Using MD, the phase-specific correlation pattern was further discovered. Compared to particle concentrations, the dissolved PAH concentrations are far more dispersed (**Figure 3.4 B**). It was discovered that Flu, Phen, Ace, and Acen maintained substantially higher mean MDs (>3.5) than all other PAH species due to their continuously common dissolved concentrations. From all other particulate PAH species concentrations, only the particulate Nap concentration showed a very high mean MD (> 5.5).

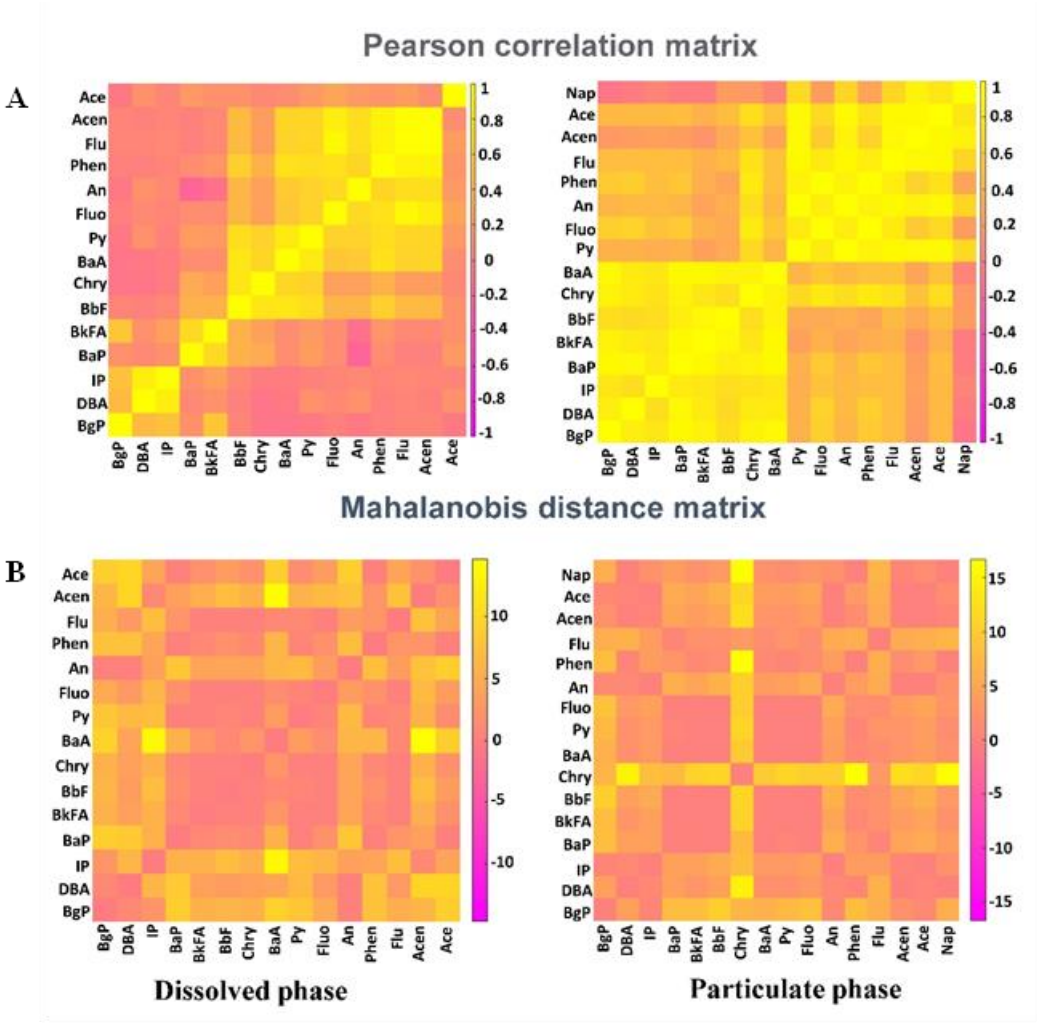


Figure 3.4 Correlation and distance matrix of all PAH species

In general, particulate PAH concentrations are much more strongly correlated than dissolved PAH concentrations, as shown in (A). Additionally, several dissolved PAH species concentrations are much more dispersed than particulate concentrations due to a higher (> 3.5) mean Mahalanobis distance from all other PAH species as shown in (B).

3.3.3 Effects of PAH phase partitioning and source location

We used the NMDS approach in order to investigate the significance of phase partitioning, which contradicts the significant seasonal variation of PAH compositions. It aims to map the seasonal PAH concentrations in a predetermined number of dimensions by solely utilising pairwise similarities that maintain the rank

order of similarities between the various PAHs. The dispersion of a plot of distance vs dissimilarity for all pairs of samples, which was condensed into a stress value, is depicted as a two-dimensional figure. It shown that the modest seasonal dispersion of particular PAHs (**Figure 3.5, Figure 3.6,** Error! Reference source not found.**)**. However, whereas particulate PAH concentrations are generally well-grouped seasonally as seen by a lower stress value of 0.1, dissolved concentrations are observed to be extensively scattered across all seasons. As can be seen, the distribution patterns of the dissolved and particulate PAH species are in sharp contrast.

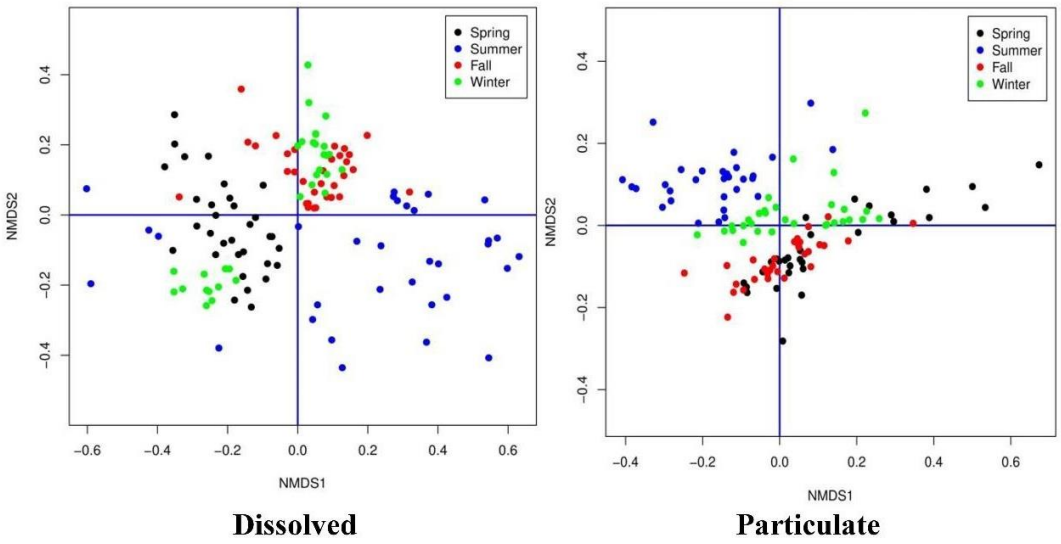


Figure 3.5 Non-Metric multidimensional scaling (NMDS) analysis

A stress value, which is a summary of the dispersion of a plot of distance vs dissimilarity for all pairs of samples, is reflected in a two-dimensional graphic. It demonstrated how the distribution patterns of the dissolved and particulate PAH species are in stark contrast.

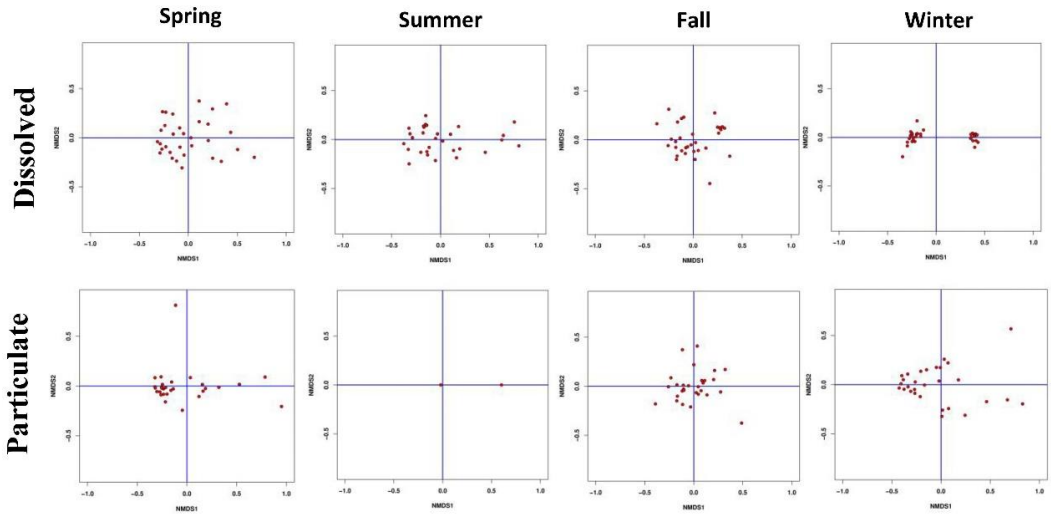


Figure 3.6 NMDS Season-Wise

Table 3.1 Season wise stress values of the NMDS

Phase	**Spring**	**Summer**	**Fall**	**Winter**	**All seasons**
Dissolved	0.180	0.115	0.188	0.045	0.131
Particulate	0.043	0.000	0.071	0.075	0.114

The stress values show how well the NMDS works. In contrast to the largely dispersed distribution of dissolved PAHs, the data demonstrated better clustering of particulate concentrations during spring, summer, autumn, and winter.

To locate the likely source of PAH contaminations, the principal component analysis (PCA) was used. **(Table 3.2)** lists the PCA outcomes. As shown, the first two axes, PC1 and PC2, accounted for 94.18% of the overall variability and explained variations in dissolved PAH concentrations of 82.29 and 11.89%, respectively. In particular, dissolved Flu and light loadings of the Acen and Ace weighed more heavily on PC1, whereas Phen alone weighed more heavily on PC2. Therefore, it was determined that PC1 was formed from petroleum, which typically produces Flu-a low molecular weight component with three aromatic rings, whereas PC2 was derived

from the dissolved Phen, with the main sources being the burning of fossil fuels, traffic, and industrial exhausts. In stark contrast, a single principal component axis, PC1, which was highly weighted by a single particulate PAH Nap, described particulate PAH variations to the tune of 99.77%. PC1 has been generated from the Nap with two rings and the lowest molecular weight of 128.1 g/mol, which has coal tar or petroleum distillation as its primary source.

Table 3.2 Principal Components Analysis (PCA) Results

	Dissolved PAHs				**Particulate PAHs**			
	PC1	**PC2**	**PC3**	**PC4**	**PC1**	**PC2**	**PC3**	**PC4**
Nap	-	-	-	-	9.96×10^{-1}	-8.52×10^{-2}	-2.61×10^{-2}	-8.30×10^{-4}
Ace	1.40×10^{-1}	6.30×10^{-2}	9.83×10^{-1}	2.97×10^{-2}	2.34×10^{-3}	8.79×10^{-3}	9.69×10^{-3}	3.69×10^{-2}
Acen	2.48×10^{-1}	-5.95×10^{-2}	-6.14×10^{-2}	9.57×10^{-1}	4.46×10^{-3}	2.87×10^{-2}	1.22×10^{-2}	5.48×10^{-2}
Flu	9.56×10^{-1}	-6.16×10^{-2}	-1.24×10^{-1}	-2.56×10^{-1}	1.57×10^{-2}	9.68×10^{-2}	-4.71×10^{-3}	1.67×10^{-1}
Phen	6.42×10^{-2}	9.87×10^{-1}	-7.86×10^{-2}	5.22×10^{-2}	5.14×10^{-2}	6.71×10^{-1}	-2.54×10^{-1}	6.48×10^{-1}
An	1.43×10^{-2}	8.90×10^{-2}	8.64×10^{-2}	-3.37×10^{-2}	3.51×10^{-3}	7.81×10^{-2}	-2.60×10^{-4}	3.17×10^{-2}
Fluo	3.98×10^{-3}	4.41×10^{-2}	-1.17×10^{-2}	-5.44×10^{-2}	2.55×10^{-2}	1.04×10^{-1}	4.56×10^{-1}	1.15×10^{-1}
Py	3.24×10^{-5}	6.59×10^{-2}	-4.14×10^{-2}	-1.04×10^{-1}	6.62×10^{-2}	7.09×10^{-1}	2.34×10^{-1}	-6.48×10^{-1}
BaA	-2.80×10^{-4}	3.03×10^{-3}	-4.46×10^{-3}	-7.14×10^{-3}	3.74×10^{-3}	-7.70×10^{-5}	3.08×10^{-1}	9.51×10^{-2}
Chry	2.95×10^{-4}	4.57×10^{-3}	-4.24×10^{-3}	1.04×10^{-2}	1.24×10^{-2}	-1.00×10^{-1}	5.44×10^{-1}	1.62×10^{-1}
BbF	5.63×10^{-4}	2.49×10^{-3}	-1.03×10^{-3}	1.72×10^{-3}	2.59×10^{-3}	-1.02×10^{-3}	1.92×10^{-1}	8.58×10^{-2}
BkFA	3.56×10^{-4}	-2.81×10^{-3}	-1.85×10^{-3}	3.82×10^{-3}	4.01×10^{-3}	-3.55×10^{-2}	1.99×10^{-1}	6.17×10^{-2}
BaP	1.51×10^{-4}	1.28×10^{-3}	-2.90×10^{-4}	-1.94×10^{-3}	4.99×10^{-3}	8.95×10^{-3}	3.01×10^{-1}	1.55×10^{-1}
IP	3.97×10^{-4}	2.50×10^{-4}	2.75×10^{-4}	-1.53×10^{-3}	1.73×10^{-3}	2.76×10^{-2}	2.45×10^{-1}	1.46×10^{-1}
DBA	4.17×10^{-4}	4.54×10^{-5}	8.17×10^{-4}	-3.91×10^{-3}	3.78×10^{-3}	2.93×10^{-2}	2.28×10^{-1}	1.49×10^{-1}
BgP	1.21×10^{-4}	1.06×10^{-3}	-1.03×10^{-3}	3.07×10^{-3}	-1.30×10^{-4}	1.29×10^{-2}	4.53×10^{-2}	3.10×10^{-2}
Explained	82.29%	11.89%	3.81%	1.44%	99.77%	0.13%	0.07%	0.018%
Mean	3.39	5.28	18.30	15.27	14.88	0.12	0.13	0.44
Eigenvalues	695.93	100.55	32.26	12.21	2180.09	2.94	1.53	0.39

According to this, the spatiotemporal variation in PAH concentrations was mostly one-dimensional and provided more than 80% of the total data.

3.3.4 Evaluation of the ecological risks posed by dissolved and particulate PAHs

Marine habitats may be at danger from high amounts of dissolved and particulate PAHs. It became essential to evaluate the possible harm caused by certain PAHs depending on their concentrations as a result. Such ecological risk assessment also enables the PAHs to be characterised in terms of any possible threats to the environment, and consequently, to ecosystems. The risk quotient (RQ) method, first suggested by[67] and later modified by[68] by include the toxic equivalence components, has been used to quantify the ecological risk of organic chemicals. Additionally, the enhanced RQ has been employed in this work to assess the various PAHs' potential risks for both dissolved and particulate concentrations.

The level of danger provided by particular PAHs is indicated by the RQ value, which is determined as follows:

$$RQ_{MPC} = \frac{C_{PAHs}}{C_{QV(MPC)}} \qquad 3.5$$

Where C_{PAHs} are specific PAHs concentration; C_{MPC} are they at the maximum permitted concentrations (MPCs) in the medium as suggested by [67] $RQ_{MPC} \geq 1$ showing the significant risk.

The current study's findings revealed that none of the PAHs, in both the dissolved and particulate phases, represent a significant harm to ecosystems throughout the seasons (**Figure 3.7**). All the dissolved PAHs, with a few notable exceptions, exhibited extremely considerable seasonal fluctuation with values less

than 0.1, in contrast to all the particulate PAHs, which all showed RQ_{MPC} values below 0.05. In particular, RQ_{MPC} of the Flu is 0.16 in the spring, 0.42 in the summer, and 0.23 in the winter, but the of the Nap RQ_{MPC} was 0.72 in the fall. Such significant seasonal change of the RQ_{MPC} shows the need for phase partitioning and PAH monitoring.

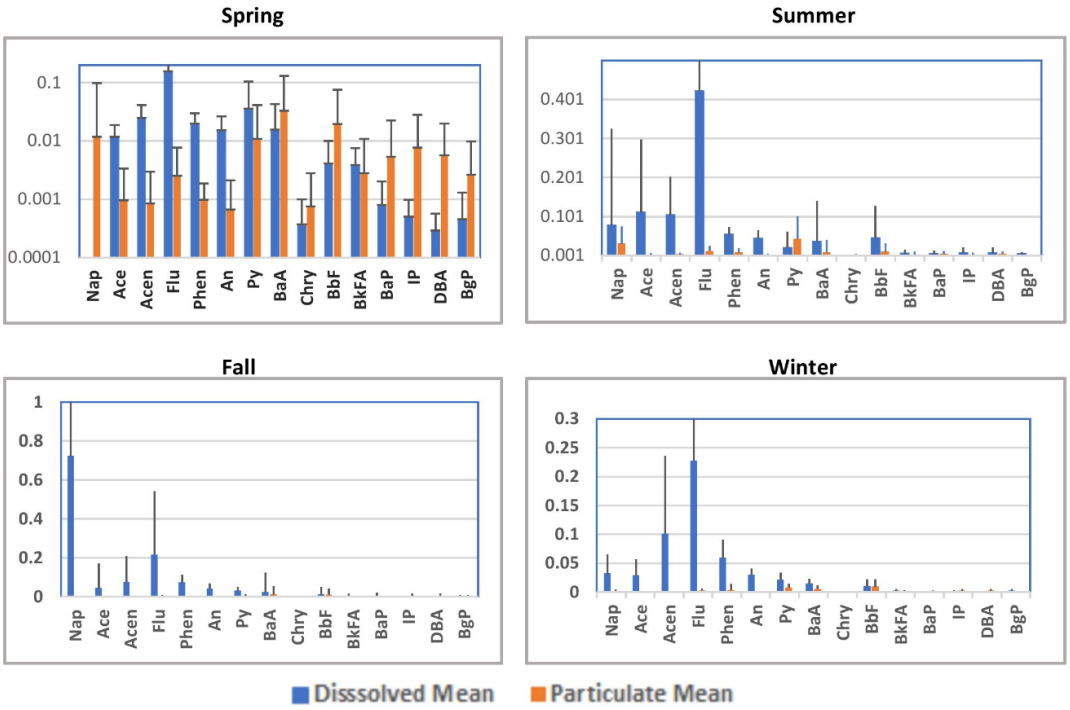


Figure 3.7 Risk quotient (RQ) variation across the four Seasons

Risk Quotient (RQ) variation across the four seasons. To determine the seasonal high-risk potential of certain PAHs, we solely computed RQ_{MPC}. All of the dissolved PAHs had RQ_{MPC} values below 0.1 with just a few notable outliers, whereas all of the particulate PAHs had RQ_{MPC} values below 0.05.

3.4 Summary

In this chapter I have conducted a spatial temporal data analysis of PAHs concentration of South China Sea (SCS) which are primarily impacted by transfers of water masses, energy, and materials between this marginal sea and the Pacific Ocean. The findings of the current research revealed significant seasonal fluctuation in PAH levels (**Figure 3.3**), nearly in agreement with the earlier studies. Furthermore, the most significant environmental variables impacting the seasonal heterogeneity and the geographical distributions of PAHs in the surface sea waters are considered to be anthropogenic activities, land- and ocean-based emissions, surface runoff, and open seawater dilution [50, 69].

2. The pyrogenic (pyrolytic) sources of PAHs include incomplete combustion of diesel fuel and engine oil, wood, coal tar, biomass from forest fires, grass fires, waste incinerators, and fossil fuels used in industrial operations and power plants. Both of these sources are related to the production of petroleum[70] At the majority of the stations under study, the sources of PAHs come primarily from petrogenic sources, with very little input from pyrogenic sources such incomplete fuel combustion in boats and car engines.

3. High levels of Nap and high concentrations of high-molecular-weight PAHs, such as Ace, Acen, Flu, Fen, An, Fluo, Py, Chry, and BkFA, were found to be significantly correlated ($p < 0.05$) (**Figure 3.4**), indicating significant secondary sources and the strong correlation of PAHs suggesting that they originated from a common source (such as wood, coal

combustion, and petroleum), which may be widely distributed in the studied location and abundant[50, 71].

4. The concentration of nap particulate remained greatest throughout all four seasons, and the MD matrix also showed that comparatively fewer PAH species had concentrations greater than 1 ng/L. The results of the NMDS clearly show the limited seasonal dispersion of PAHs and the divergent impacts of phase partitioning. Using PCA, it was possible to determine the main sources of phase-wise PAH contamination, which were either petroleum distillation or coal tar.

In both the dissolved and particle phases, RQ values revealed that no PAH poses a significant threat to ecosystem health across the seasons; only the particulate Nap exhibited a comparatively high value of RQ_{MPC} of 0.72 in the fall. The spatiotemporal fluctuations of PAHs may be further influenced by several important processes, such as air-sea exchange and deep-sea burial. To improve this current study and analyse regional and global source-sink dynamics of the PAHs in the future, these must be looked at.

"Spatiotemporal data analysis is the key to understanding the dynamic patterns and processes that shape our world" …… Michael Goodchild

Chapter 4

4. Spatiotemporal data modelling and analysis-II

Graphical View

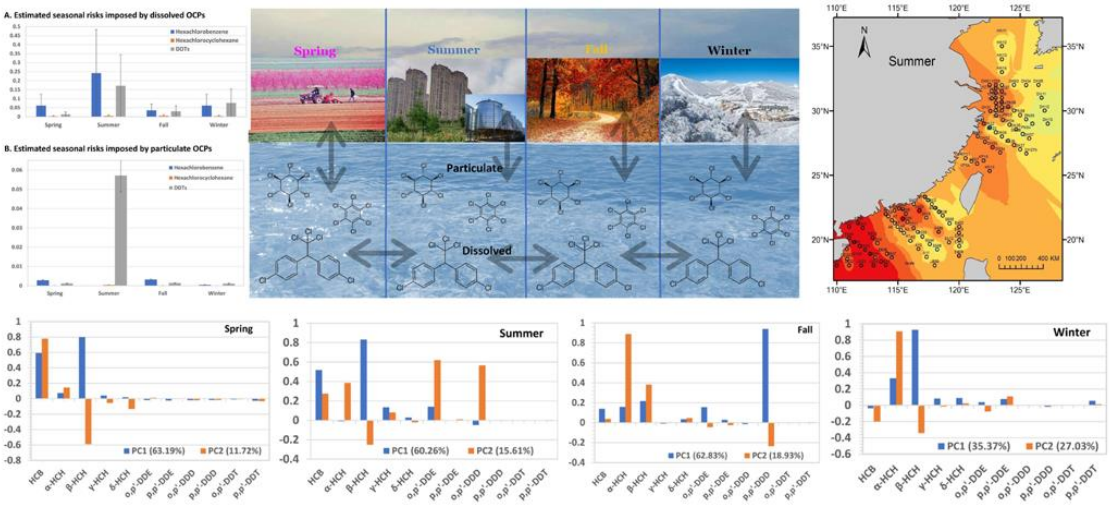


Figure 4.1 Graphical Abstract

Wang, C., Feng, L., **Thakuri, B.**, Chakraborty, A. (2022). Ecological risk assessment of organochlorine pesticide mixture in South China Sea and East China Sea under the effects of seasonal changes and phase-partitioning. *Marine Pollution Bulletin*, *185*, 114329.

4.1 Introduction

Organochlorine pesticides (OCPs) are persistent organic pollutants (POPs) that are commonly utilised as broad-spectrum insecticides for efficient agricultural pest control [72]. They are highly poisonous, resistant to biodegradation, quickly accumulated in food chains, and capable of biomagnification, making them the most dangerous agents with a severe influence on the environment and ecosystems [73, 74]. OCPs have drawn significant global attention due to their widespread use and known negative effects on human health, including cancer, reproductive defects, and endocrine and immunological toxicities [75]. As a result, OCPs are illegal in the majority of developed nations in the Northern Hemisphere. However, several Southeast Asian nations still employ them [76-79]. OCPs continue to pose a severe worldwide threat because to their long-distance transport capabilities from the sources through recurrent evaporation-condensation processes, even though their current usage are constrained to certain nations and regions. OCPs are easily bound by suspended particulate matter (SPM) in water and air and often redistributed by absorption on solid particles due to their high values of octanol/water partition coefficient and poor water solubility[80, 81] . OCPs precipitate to the bottom of water after being absorbed by allochthonous and autochthonous particles. OCPs that are collected from several sources, including neighbouring agricultural sites, surface runoff from catchment regions, fall with precipitation, atmospheric circulation, and movement by ocean currents, are frequently a sink for marine and aquatic ecosystems[69, 82, 83] . As a result, it continues to be important to monitor and better regulate the distribution, incidence, and ecological threats caused by marine OCP. The most severely harmed marine habitats are those that border continents and oceans, and

these ecosystems have been identified as being crucial to the source-sink dynamics of POPs. With a combined area of about 4.7 million km2, China’s marginal seas, which include the Bohai Sea, Huanghai Sea, East China Sea, and South China Sea, are particularly affected by eutrophication, overfishing, excessive land reclamation, adjacent land use changes, and climate change. As a result, the seas' ability to support human well-being is declining [84-86]. The marginal seas of China provide significant sea-based commercial links between the Eurasian, Pacific, and Indian-Australian areas, while the neighbouring land regions are currently seeing rapid population growth. According to current reports, during the past 20 years, China’s marginal waters have received almost 5000 tonnes of DDT releases[87] . These OCPs may be carried with the particles to the deep sea and buried there or may be partly destroyed by plankton. Along with serving as a substantial global OCP sink, it also releases a sizeable quantity of OCPs through processes including air-water exchange and volatilization[88] . Estimating occurrences, variability, and risk assessments become crucial for sustainable management of OCPs and its impact because the marginal seas of China are heavily impacted by historically largest and most widespread use of OCPs in China and Southeast Asia and that is largely controlled by seasonally alternating East Asia monsoon.

The classic methods of risk assessment focus on the risk estimations of a particular kind of chemical. An environment that is frequently exposed to chemical mixtures rather than a single item may have more danger than this evaluation indicates [89, 90]. A chemical with a concentration below the no observed effect can nonetheless have a combined impact when it is taken into account in a combination risk calculation. Consequently, ecological risk evaluations of co-exposed

contaminants based on mixture risk models (MRM) can offer more accurate estimates of a variety of dangerous compounds [91]. This method was used to estimate the risk associated with OCP mixtures in the SCS and ECS.

4.2 OCPs species

The 11 OCP species that were taken into account in this study were divided into three broad groups based on their chemical similarities: DDTs (o, p'-DDE; p, p'-DDE; o, p'-DDD; p, p'-DDD; o, p'-DDT; p, p'-DDT), hexachlorocyclohexane (HCH (α-HCH; β-HCH; γ-HCH; δ-HCH), and hexachlorobenzene (HCB)(). has a list of their fundamental chemical characteristics.

4.3 Methods and materials

4.3.1 Sample collection and the study region

The scientific research ship "Dongfanghong-2" collected surface water samples from each established station (with a depth of no more than 2m) throughout the SCS and ECS continental shelf borders from 2009 to 2011. This fieldwork was done in the spring of 2011 (April-June), the summer of 2009 (July-August), the fall of 2010 (October-December 2010), and the winter of 2009-2010. (**Figure 4.2**). At each station, a submersible pump was utilised to transfer the water into a 50-liter steel barrel. Then, particle was separated using a peristaltic pump and a filtration system (particles with size > 0.75 μm) and dissolved phase. To remove dissolved OCPs from the filtered water, two parallel solid-phase extraction (SPE) processes with prior cleaning were used. Prior to enriching the water sample, an activated SPE column containing 10 mL of methanol was eluted with 10 mL of ultrapure water (resistivity of 18.25 MΩ·cm). Finally, pre-fired aluminium foil paper was used to wrap the granular

phase and dissolved phase SPE column samples and seal them within a small bag. The SPE column samples from the granular phase and dissolved phase were then packed in a small bag and covered in pre-fired aluminium foil sheets. At -4 °C, it was frozen and kept.

Table 4.1 Basic chemical properties of 11 OCPs grouped into three primary classes

OCP class	Chemicals	CAS	Molecular Weight [g/mol]	Log K_{ow}	Solubility (25°C) [mg/L]
Hexachlorobenzene	HCB	118-74-1	389.3	5.40	4.7×10^{-3}
Hexachlorocyclohexane	α-HCH	319-84-6	290.8	3.80	2.0
	β-HCH	319-85-7	290.8	3.78	0.24
	γ-HCH	59-89-9	290.8	3.72	7.3
	δ-HCH	319-86-8	290.8	4.14	31.4
DDTs[#]	o, p'-DDE	53-19-0	320	7.00	0.14
	p, p'-DDE	72-54-8	320	6.51	0.04
	o, p'-DDD	3424-82-6	320	5.87	0.10
	p, p'-DDD	72-55-9	320	6.02	0.16
	o, p'-DDT	789-02-6	320	6.70	7.68×10^{-2}
	p, p'-DDT	50-29-3	320	6.91	4.96×10^{-3}

Source: https://pubchem.ncbi.nlm.nih.gov; https://comptox.epa.gov; https://www.drugbank.ca

DDE: 1,1-dichloro-2,2-bis(p-chlorophenyl) ethylene; DDD: 1,1-dichloro-2,2-bis(4-chlorophenyl) ethane; DDT: 1,1,1-trichloro-2,2-bis(4-chlorophenyl) e

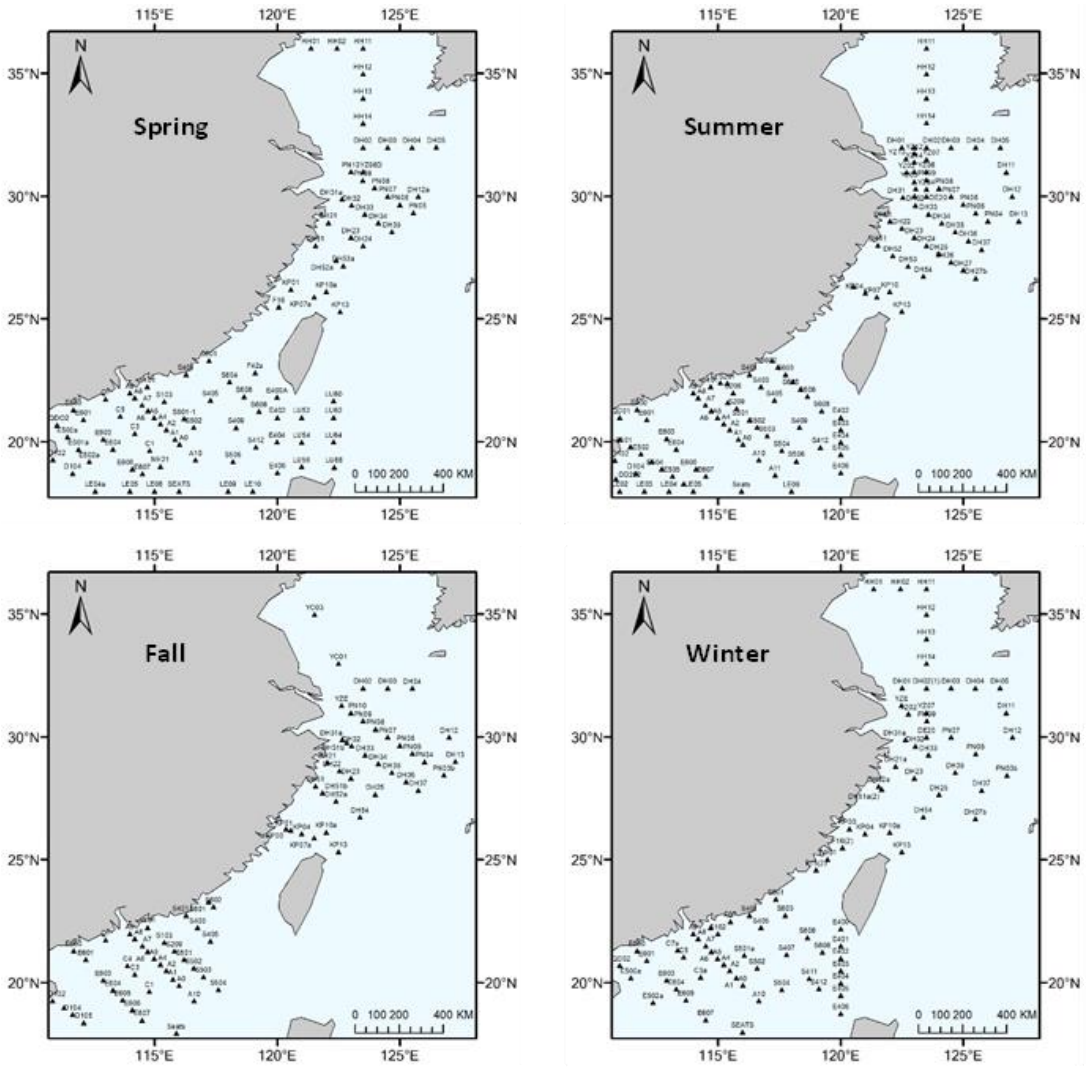


Figure 4.2 Sampling location along the South and East China Seas

~4.7 million km^2 of the South and East China Seas were sampled at the following locations: Spring 92, Summer 116, Fall 78, and Winter 85. There were 30 total sites where the dissolved and particulate OCPs overlapped.

4.4 Data analysis

For the 11 OCP species, chemical concentrations were found. Four seasons of sample sites were used to analyse the measured data's seasonal and regional variance. The number of sampling locations (spring: 92; summer: 116; fall: 78; winter: 85) and location-specific water depth (18-4795; 15-1051; 16-3848; 20-4180m), site-specific surface temperature (24.913.77; 28.561.58; 21.075.62; 19.095.21 °C), and overall water quality (salinity: 33.39±1.09; 31.93±2.72; 33.45±1.47; 33.44±1.37%, and SPM: 5.21-34.23; 4.46-66.89; 7.75-93.77; 0.96-130.04 mg/L). In all four seasons, there were just 30 counts, which is a relatively low number of overlapping locations. To evaluate the phase-partitioning effects, the dissolved and particulate phases were separated.

4.4.1 K-means clustering

A form of unsupervised learning is K-Means clustering. Finding groups in data is the major objective of this technique, and K stands for the number of groups. Each data point is sorted iteratively into one of the K groups according to how comparable its features are. Starting with K centroids that are randomly chosen from the dataset as starting estimates, the K-Means method is used. The method repeats two operations: allocating data points and updating Centroids.

By dividing the sampling sites according on the concentration profiles of 11 OCP species, the K-means clustering technique was used to analyse site-to-site variations over the four seasons. Each site belonged to one of the clusters with the closest mean once clustering was complete (i.e., cluster centroid). Expectation-Maximization strategy was used with the K-means clustering technique [92, 93]. The

data points are assigned to the nearest cluster in the expectation stage, and the cluster centroid is calculated in the maximising step. The data points were assigned to clusters in this procedure such that the sum of the squared distances between the data points and the centroid would be as small as possible. Within the clusters, which represented more comparable data points, less variance was retained.

Out of 100 initializations that generate the best total sum of distances, we used K-means clustering in MATLAB version R2021a with the input of 2 clusters and picked the best initial centroid arrangement. We chose a good distance measure by using the approach of trial and error, which minimises the distance from the centroid and displays the lowest level of overlapping dots[94, 95]. The L1 distance, or the sum of absolute differences (Eq. 4.1), was shown to produce the best clustering results:

$$\boldsymbol{d(x,c) = \sum_{j=1}^{n} |x_j - c_j|} \qquad \textit{4.1}$$

where x_j stood for a data point and c_j for the centroid.

4.4.2 Principal Component Analysis (PCA)

To ascertain the connections between various sources and OCP concentrations, PCA was used. It breaks down the complicated data into a small number of dimensions known as principle components (PCs)[96, 97]. There is no correlation between any two PCs. The variability associated with the collection of independent PCs, which is quantified by the share of total variance that each PC accounts for in Equation (4.2), determines the characteristics of PCA.

$$\Delta_j = \frac{\lambda_j}{tr(A)} \qquad 4.2$$

where λ was taken to be the eigenvalue of the matrix A and tr(A) stood for the trace of the covariance matrix.

Using the built-in PCA function in MATLAB version R2021a, which essentially employs the singular value decomposition technique, we implemented the PCA. It returns the principal component coefficient matrix, which is an 11 by 11 matrix with one principal component coefficient in each column. PCs are sorted according to component variance in descending order.

4.4.3 Hausdorff Distance Measure (HDM)

When two data sets that show how far apart two things are from one another are compared, the Hausdorff distance is calculated [98, 99]. It is the greatest possible distance between any two points in one set and their closest neighbours in the other set. If the distance between any two points on two sets is "not too far", then two sets are said to be "similar or near". If there are two finite point sets, $A = \{a_1, a_2, ..., a_p\}$ and $B = \{b_1, b_2, ..., b_q\}$, which stand in for two sets of observations, the HDM is defined by Equation (4.3) as follows:

$$\boldsymbol{d_H(A,B) = max\{sup_{x\in A} inf_{y\in B} d(x,y), sup_{y\in B} inf_{x\in A} d(x,y)\}} \qquad 4.3$$

where the distance metric, d, was frequently seen as $d(x,y) = \|x-y\|$ and $\|.\|$ served as an example of the Euclidean norm. We used this metric to assess how different the overlapping sites were from one another throughout all seasons and in both the dissolved and particulate phases. It was carried out in MATLAB R2021a.

4.4.4 Mixture risk model (MRM)

For the estimates of spatiotemporal ecological risk, a two-tier MRM was taken into account. To investigate the impact of phase-partitioning between the dissolved and particulate forms of 11 OCPs, two separate evaluations were conducted. RQ was computed as a ratio between the recorded OCP concentration and the predicted no-effect concertation (PNEC), also known as the non-observed effect concentration (NOEC), in the first tier using concentration addition techniques at each location. The combination RQs were then determined by adding together all the species that belonged to the three main groups of OCPs, HCB, HCH, and DDTs. Three example trophic levels were added in order to take into account the ecological consequences by substituting the NOEC/PNEC with the half-effect concentration (EC_{50}) of algae and *Daphnia* and the half-lethal concentration (LC_{50}) of fish, adjusted by suitable evaluation factors extracted from [87](. The U.S. EPA's ECOSAR (Ecological Structure Activity Relationships) database contains estimates of EC_{50} and LC_{50}. Finally, Equation (4) was used to estimate the mixture RQ:

$$RQ_i = \sum_j \frac{MC_{ij}}{\min(EC_{50j,algae}, EC_{50\,j,Daphnia}, LC_{50\,j,fish}) \times (1/AF)} \quad (4.4)$$

where i=HCB, HCH, and DDTs; MC_{ij} was measured concentration of jth species of the OCP class i and AF was assessment factor. According to different ranges of RQ values, risk levels were graded as follows: low risk: 0.01<RQ<0.1; medium risk: 0.1<RQ<1; and high risk: RQ>1.

The second layer of the MRM only operated when RQ > 1, and RQSTU was determined by adding the toxic units of the most vulnerable organism group for each sample trophic level, as indicated in Eq. (4.5):

$$RQ_{i,STU} = max\left(\sum_j \frac{MC_{ij}}{EC_{50j,algae}}, \sum_j \frac{MC_{ij}}{EC_{50j,Daphnia}}, \sum_j \frac{MC_{ij}}{EC_{50j,fish}}\right) \times AF \quad 4.5$$

where LC_{50} stood for half deadly concentration (mg/L) for the fish, EC_{50} stood for half impact concentration (mg/L) for the algae and Daphnia, and AF was set at 100.

4.5 Results

4.5.1 Composition and geographic spread of the OCP phase

In two separate stages, dissolved and particulate OCPs, seasonal sampling and concentration determination of 11 OCP species in SCS and ECS were conducted. The average particulate concentrations were 0.022, 0.089, and 0.322 ng/L, whereas the total mean dissolved concertation of HCB, HCH, and DDTs across seasons was 1.001, 2.531, and 1.168 ng/L, respectively (**Table 4.2**). With significant seasonal fluctuations, dissolved OCPs have consistently kept much greater concentrations than the particulate phases for all three groups. In contrast to HCB and DDTs, dissolved HCH maintained a very high concentration throughout all the seasons. Except for summer DDTs, all three groups' dissolved OCP concentrations stayed at least 10 times greater than their particulate counterparts in all seasons. Seasonal and phase-based sharp contrasts in dispersion have been seen. Particularly, particulate DDT concentrations in the summer were more than 10 times greater (1.196 ng/L) than in any of the other seasons and were generally higher than dissolved DDT

concentrations in the spring, fall, and winter, with the exception of summer (2.386 ng/L). With the exception of the fall, when the concentration of HCB is lower than that of DDTs, dissolved OCPs consistently kept the order HCHs>HCB>DDTs throughout all the seasons. Contrary to this component order, particulate concentrations varied only slightly across all OCP groups and seasons, with summer elevated DDT sticking out as an exceptional exception. Additionally, we observed that the mean prevalence of dissolved HCHs increased with the seasons: winter 2.24 ng/L (1.98-2.51 ng/L), autumn 3.26 ng/L (3.04-3.48 ng/L), summer 1.708 ng/L (95% CI: 1.46-1.96 ng/L), and summer 2.91 ng/L (2.20-3.62 ng/L). However, compared to dissolved forms, the mean concentrations of all OCPs in the particulate phase remained much lower, while the mean concertation of HCHs in the particulate phase was higher in the spring (0.033 ng/L, 0.025-0.041 ng/L), summer (0.247 ng/L, 0.196-0.298 ng/L), and fall (0.067 ng/L, 0.056-0.078 ng/L). DDT concentrations remained greater due to winter particulates

Table 4.2 Seasonal and phase-wise occurrences and distribution of major OCPs

Seasonal & Phase-separation effects	**Spring**		**Summer**		**Fall**		**Winter**	
	Dissolved	**Particulate**	**Dissolved**	**Particulate**	**Dissolved**	**Particulate**	**Dissolved**	**Particulate**
Hexachlorobenzene (HCB)								
Mean concentration (ng/L)	0.619	0.028	2.417	-	0.351	0.032	0.617	0.005
95% CI	0.464-0.774	0.024-0.033	2.018-2.816	-	0.293-0.409	0.027-0.037	0.502-0.732	0.002-0.008
Mixture Risk Quotient (MRQ)	0.062±0.076	0.003±0.002	0.242±0.219	-	0.035±0.026	0.003±0.002	0.062±0.054	0.001±0.001
Hexachlorocyclohexane ($\sum_{\alpha,\beta,\gamma,\delta} HCH$)								
Mean concentration (ng/L)	1.708	0.033	2.910	0.247	3.261	0.067	2.243	0.009
95% CI	1.457-1.959	0.025-0.041	2.198-3.622	0.196-0.298	3.041-3.481	0.056-0.078	1.979-2.506	0.004-0.014
Mixture Risk Quotient (MRQ)	0.003±0.002	0.000±0.000	0.005±0.007	0.000±0.000	0.005±0.002	0.000±0.000	0.004±0.002	0.000±0.000
DDTs ($\sum DDE, DDD, DDT$)								
Mean concentration (ng/L)	0.330	0.026	2.386	1.196	1.112	0.029	0.843	0.036
95% CI	0.236-0.423	0.022-0.030	1.747-3.025	0.874-1.518	0.814-1.411	0.024-0.034	0.527-1.160	0.023-0.048
Mixture Risk Quotient (MRQ)	0.013±0.022	0.001±0.001	0.172±0.283	0.057±0.089	0.030±0.043	0.001±0.001	0.077±0.075	0.001±0.003
Major contributor (Along PC1)	β-HCH (80%)	α-HCH (85%)	β-HCH (83%)	o, p'-DDT (71%)	p, p'-DDD (93%)	α-HCH (81%)	β-HCH (93%)	p, p'-DDT (93%)

With the appropriate number of groups and distance measure, we used K-means clustering to investigate site-to-site spatial and seasonal variations (**Figure 4.3**).

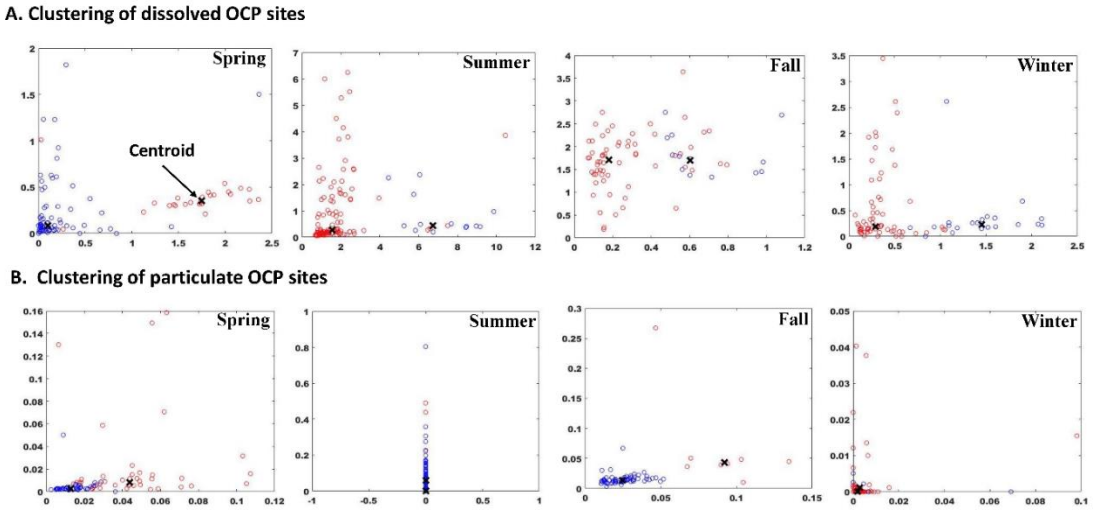


Figure 4.3 K-means clustering under L1 distance matric

(A) dissolved phase and (B) particulate phase. It indicated large seasonal variation and phase partitioning effects on the OCP distributions.

Interestingly, it revealed significantly less spatial variability of OCPs in both stages and all seasons. Since the centroid of two possible clusters were either extremely close to one another or exhibited layered collinearity of datapoints, no substantial grouping in particulate OCP sites occurred during the summer and winter seasons in particular. In comparison, dissolved OCP locations displayed superior clustering results. With total sums from the respective centre of cluster-1 and cluster-2 being (spring: 35.42, 94.81) and (fall: 93.09, 37.81), respectively, in the dissolved phase, spring and fall showed the almost non-overlapping clusters, showing greater resemblance among the sites.

We used HDM to measure the differences between the 30 overlapping locations over the four seasons. Overall, it revealed low HDMs that were mostly identical across all four seasons in both stages, with very few locations standing out from the rest (**Figure 4.4**). Thus, it suggests less geographic variation in line with the findings of K-means clustering. In comparison to particulate phase, mean HDM in the dissolved phase reliably remained higher throughout all the seasons, indicating larger spatial variance in this phase.

Hausdorff Distance map that measures the degree of dissimilarity among 30 overlapping sites across seasons and between the two phases. Overall, it indicated lesser spatial heterogeneity. Relatively, the dissolved mean Hausdorff measure consistently remained greater across all the seasons than particulate phase, representing higher spatial variation in this phase. The greatest mean HDM and standard deviation were found to be maintained during the summer season (dissolved: 2.1±2.16; particulate: 1.0±1.46), with approximately the same range of HDM variation. With the exception of the summer season, dissolved phase HDM long-range variance has been found to be at least three times greater than that of particulate phase.

In conclusion, the clustering and HDM findings unequivocally show that the OCP distributions are affected by phase splitting and significant seasonal variance. While spatial variance in both phases has persisted at low levels, able to sustain a nearly homogeneous environment—an known typical feature of marginal seas—this has been the case in both phases.

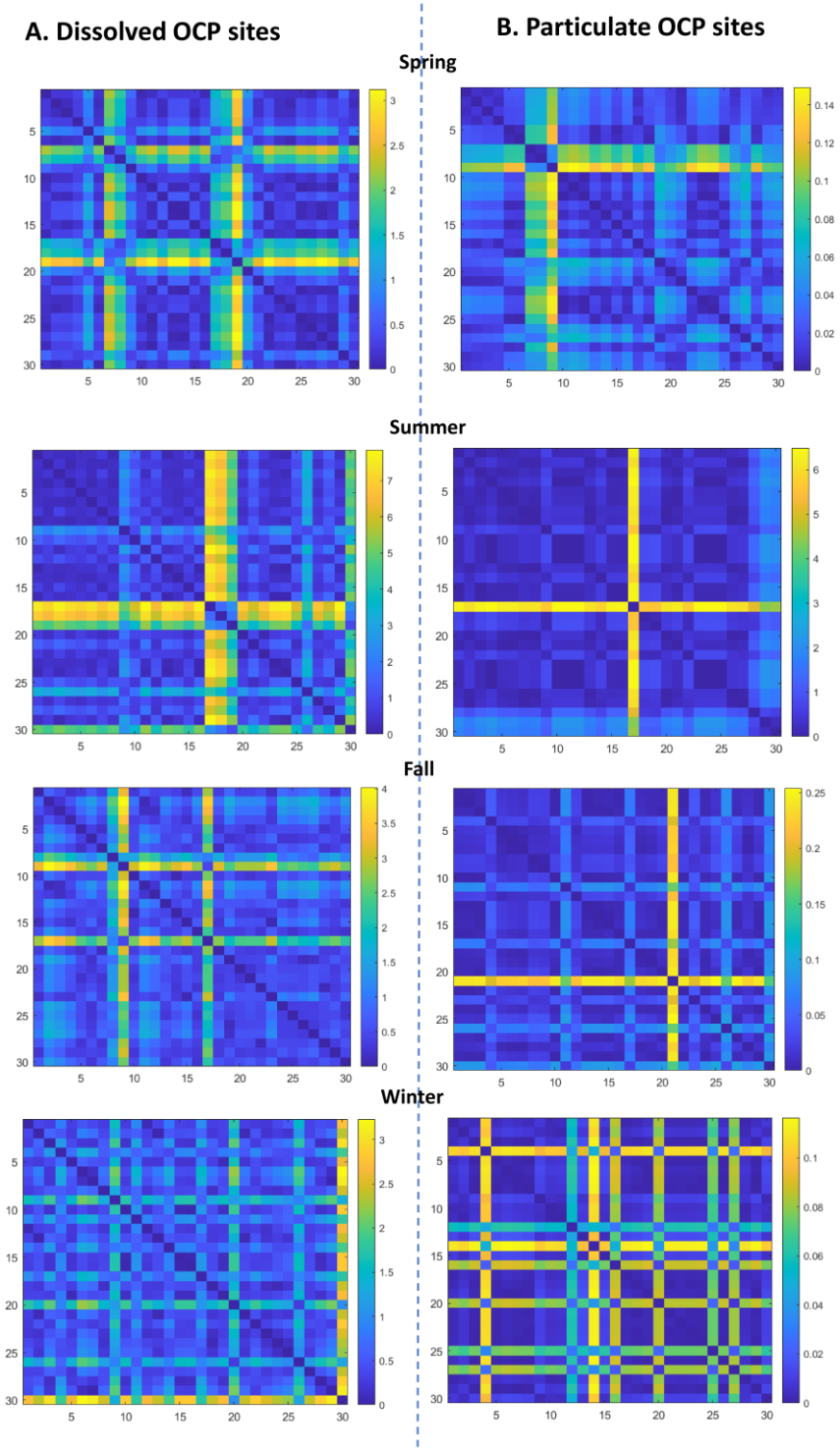


Figure 4.4 Hausdorff Distance Map

Using a Hausdorff Distance map, 30 overlapping locations are compared for similarity between seasons and between the two phases. It showed less spatial heterogeneity overall. In comparison to particulate phase, the dissolved mean Hausdorff measure consistently stayed higher during all the seasons, indicating larger spatial variance in this phase.

4.5.2 Different prevalent OCP species have distinct impacts.

We were able to locate possible OCP sources in the SCS and ECS by using PCA to clarify prominent contributors and their seasonal and phase-wise variation. The findings indicated that the first four PCs accounted for more than 80% of all OCP variance in both stages and throughout all seasons. By performing the PCA separately for each season and phase, seasonal and phase-wise variance was caught in the analysis, producing domination patterns of OCPs with the underlying effects of seasons and phase partitioning (**Figure 4.5**).

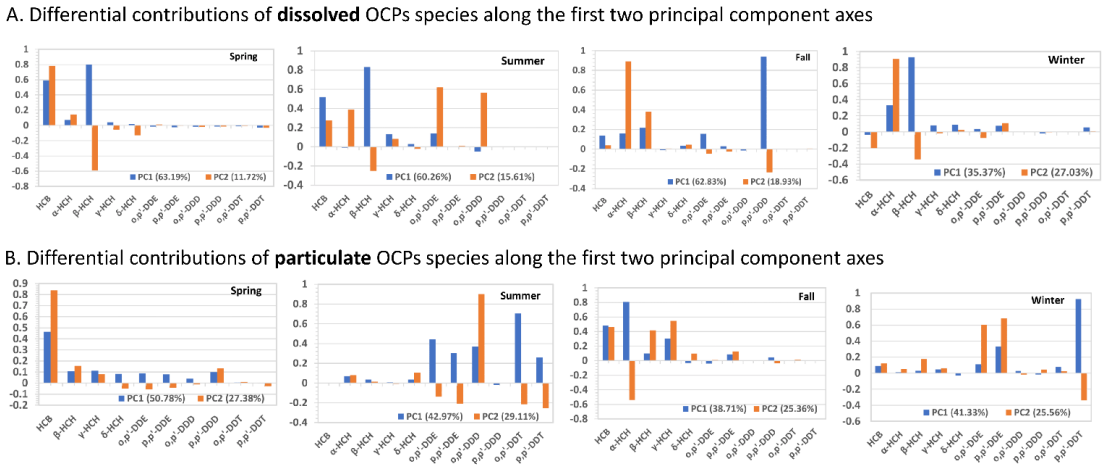


Figure 4.5 Principal component Analysis

Principal component analysis of (A) the dissolved phase and (B)the particulate phase, respectively, demonstrating the proportional positive and negative loadings of each individual OCP species in PC1 and PC2. More than 80% of all OCP concentration changes were described by the first four PCs. The fact that any of the HCH, DDT species, and HCB favourably loaded either of the axis shows that there are numerous sources of OCP contamination in the SCS and ECS.

In the spring, summer, autumn, and winter, respectively, principal component 1 (PC1) explained 63.19, 60.26, 62.83, and 35.37% of the overall variation in dissolved OCP. In spring, summer, and winter, it was reliably loaded by the HCH component with the greatest input (>0.8). In contrast to all other seasons, the contributions of p, p'-DDD contributed >0.9 in the fall, and almost equal contributions (~0.2) of different HCH species and HCB were seen. Across all the seasons, β-HCH made the largest contribution of all the HCH species. As opposed to this, PC1 of particulate OCPs explained 50.78, 42.97, 38.71, and 41.33% of the overall variation in the spring, summer, fall, and winter, respectively. There was no OCP species or class that consistently dominated. Instead, it was found that DDTs made very significant contributions in the summer and winter, while HCB and HCH species predominated in the spring and autumn. It's interesting to note that high particulate α-HCH input (>0.8) has been observed in the spring and fall.

The difference between various OCP species contributions and consequently various pollution sources is described by principal component 2 (PC2). In the spring, summer, autumn, and winter, it described 11.72, 15.61, 18.93, and 27.03% of the overall variation in dissolved OCP, respectively. Across seasons, there were generally fewer positive loadings by different species, but the inventory was unusually enlarged in the summer. Furthermore, it was uncommon for a single species to reliably make significant contributions across seasons. There were very clear high summer loadings of PC2 by o, p'-DDE, o, p'-DDD, α -HCH, and HCB. PC2 accounted for 27.38, 29.11, 25.36, and 25.56% of the overall variance in the particulate phase. Different species were observed to positively load PC2 in stark contrast to the dissolved phase. Particulate HCB (0.84) and β-HCH (0.16), p, p'-DDD (0.14) highly contributed in

spring; γ-HCH (0.55), HCB (0.46), β-HCH (0.41), p, p'-DDE (0.13) in fall; o, p'-DDD (0.9) in summer; and p, p'-DDE (0.68), o, p'-DDE (0.6), β-HCH (0.18), HCB (0.12) in the winter. Multiple sources of OCPs are indicated by the fact that various multiple species in each of the three OCP groups demonstrated strong positive contributions across seasons.

4.5.3 Evaluation of the spatiotemporal ecological risk

The spatial ecological risk evaluations used the MRM. Algae, daphnia, and fish are three exemplary trophic layers included in this. RQ values were used to rate the risk levels: low risk (-0.01<RQ<0.1); middle risk (-0.1<RQ<1); and high risk (RQ>1). The second layer of the MRM only functioned when RQ > 1, and the total of toxic units of the most vulnerable creature for each sample trophic level was used to compute RQ_{STU}. Risk distribution maps were then created for each season and phase, taking into account the found seasonal and period petitioning impacts. Except for the summer season, (**Figure 4.6**) demonstrated that none of the OCP groups caused any high-risk zones during any of the seasons or phases. Only one location for dissolved HCB and two sites for dissolved DDTs had the RQ greater than one in the summer. Therefore, it demonstrated that the amounts of both dissolved and particulate OCP were much lower than the no-effect limits for causing the elevated risk. In addition, it was observed that the majority of particulate OCPs did not represent a risk throughout the seasons. In particular, particulate DDTs revealed low-risk at 52.29% of the locations and medium risk at 14.68% of sites during the summer. In contrast, there is little danger from particulate HCB (3.37% in spring and 3.95% in fall). In stark contrast, the danger presented by dissolved OCPs was low to moderate throughout the year. In the summer, dissolved HCB (81.90%) and DDTs (34.25%) were found to

pose the highest medium risk. Low level risk primarily caused by HCB and DDTs; in the fall and winter, HCH presented a much lower risk (~10 times). At 12% of locations during the summer, dissolved HCH showed a moderate degree of risk. RQ_{STU} computed at uncommon sites in the second tier of the MRM (**Table 4.3**) revealed RQ_{HCB} and RQ_{DDT} >1. It revealed that RQDDT accounted for *daphnia* (0.25) and RQ_{HCB} for algae (1.046).

In conclusion, it was found that the spatial ecological risk evaluation in the SCS and ECS was significantly influenced by seasonal and phase partitioning effects. The RQ computation revealed seasonal low-to-medium level hazards that were mainly caused by dissolved HCB and DDTs. Low-level DDT-posed hazards in the particle phase were only seen during the summer. Only a small number of locations showed elevated risk from HCB and DDTs, which are sensitive to algae and daphnia, respectively.

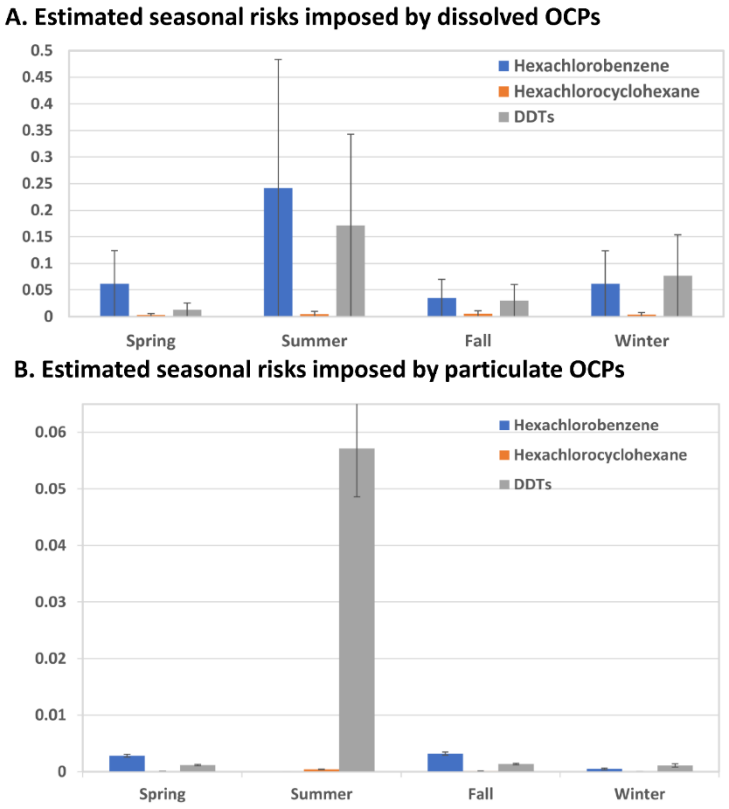


Figure 4.6 Mixture risk Model estimated ecological risk posed by major OCP classes

Mixture risk model estimated ecological risk posed by major OCP classes: (A) dissolved OCPs and (B) particulate OCPs. It showed seasonal variation and the phase partition effects on the estimates. Dissolved HCB and DDTs posed higher risk across all the seasons. Whereas in particulate phase DDT-posed significant risk was noted in summer.

Table 4.3 Ecological mixture risk assessment across seasons and phase in South China Sea and East China Sea

Season	**Risk levels**	**RQ Summary (Dissolved)**			**RQ Summary (Particulate)**		
		HCB	**HCH**	**DDT**	**HCB**	**HCB**	**HCH**
Spring	No. High risk (>1)	0.000	0.000	0.000	0.000	0.000	0.000
	% Low risk (0.01-0.1)	36.667	0.000	29.688	3.371	0.000	0.000
	%Medium risk (0.1-1)	27.778	0.000	1.563	0.000	0.000	0.000
summer	No. High risk (>1)	1 (RQ_{algae}=1.046) *	0.000	2 ($RQ_{daphania}$=0.245) *	0.000	0.000	0.000
	% Low risk (0.01-0.1)	17.241	12.069	32.877	0.000	0.000	52.294
	% Medium risk (0.1-1)	81.897	0.000	34.247	0.000	0.000	14.679
Fall	No. High risk (>1)	0.000	0.000	0.000	0.000	0.000	0.000
	% Low risk (0.01-0.1)	89.744	1.282	27.632	3.947	0.000	0.000
	% Medium risk (0.1-1)	1.282	0.000	13.158	0.000	0.000	0.000
Winter	No. High risk (>1)	0.000	0.000	0.000	0.000	0.000	0.000
	% Low risk (0.01-0.1)	72.941	1.176	65.714	0.000	0.000	2.532
	% Medium risk (0.1-1)	24.706	0.000	22.857	0.000	0.000	0.000

* when RQ>1, RQ_{STU} was calculated by taking sum of toxic units of the most sensitive organism group.

4.6 Discussion

4.6.1 OCPs' spatiotemporal heterogeneity

The marginal seas, which are located at the nexus of terrigenous and oceanic habitats, are busy year-round and are essential to sustaining the source-sink dynamics of OCPs. At the international level, the marginal China Sea, which is bordered by Southeast Asian Countries, plays a significant part as an OCP source-sink zone. Although OCP use has been outlawed in China for more than 30 years, the country has traditionally made extensive use of HCHs and DDTs, accounting for more than 20% of global usage in the decades prior to their outlaw [100-102]. OCPs that were caused by human activity are still widely distributed in the earth, water, air, sediments, and living things. This research demonstrated that, with the exception of the autumn and winter, mean concentrations of dissolved HCB, HCHs, and DDTs in the summer (>2.3 ng/L) were significantly greater than those in other seasons. The mean amounts of dissolved DDT were consistently greater in the summer (2.386 ng/L) and winter (0.843 ng/L). When compared to the published amounts of HCB (0.0021-0.0061 ng/L), DDTs (0.002 ng/L), and HCHs (0.09-0.627 ng/L) in the subtropical North Atlantic Ocean, these observed concentrations in the SCS and ECS are still significantly higher (at least 10 times) [103]. Additionally, it is considerably greater than the HCH and DDT concentrations that have been observed in the open Pacific Ocean[69]. The monsoons dominate the climate in the China Sea's peripheral region, which is equatorial. Typhoons are common in the summer, and south westerly winds dominate during the rainy season. Winter breezes, however, come from the northeast. The annual precipitation is between 2000 and 3000 millimetres[104] . The spread of OCPs is substantially impacted by this warm weather. The current research found that

all OCP amounts in the SCS and ECS had extremely high seasonal fluctuation. The study of phase separation between dissolved and particulate form increases this diversity even further. It can be ascribed to a variety of combinations of the OCPs' seasonal changes in transport paths from different sources. Surface discharge, volatilization, biological mechanisms, diffusion, and ocean currents are examples of possible transport pathways [46, 105, 106]. Due to comparatively lower concentrations of prohibited OCPs in the oceans, a high-to-low concertation gradient can be anticipated at the land-sea boundary. Terrigenous OCPs may be diffusely transported into oceans as a result of this concertation gradient. Surface runoff has the potential to be a major OCP transport pathway during lengthy and strong monsoons in Southeast Asia [107]. A high-temperature-enhanced volatilization of OCPs, which results in their diffusion from soil to atmosphere and ultimate deposition into oceans, may be the cause of the relatively much higher summer OCP content found in the SCS and ECS. In addition, the summer's typically low algae density can impede the degradation process, increasing the content of OCP.

The distribution and destiny variations of various OCPs were significantly influenced by phase separation between dissolved and particulate forms. The SCS and ECS were significantly affected by such impacts. All of the dissolved OCPs in the current research revealed a comparatively greater content, whereas the particulate DDTs demonstrated their considerable amount in comparison to all other species. Since it could not be identified at a significant level when measured in combination, it indicated the significance of phase separation. The ecological risk maps showed this phase difference as well. It stood out from all other OCPs due to low water solubility of DDTs and its strong attraction for particulates in the water column, which affected

how widely it was distributed[108]. For instance, DDT linked with SPM found to be 90% in the Peral River Estuary, 40% in the Yangtse River, China, and up to 54% in the coastal seas of Singapore [109, 110]. This research discovered that the mean particulate DDT levels in the summer were greater than those in the other seasons and similar to their dissolved content. It also discovered that these levels were observable in the winter and autumn. In comparison, HCH isomers (α, β, γ, δ) were found to have a minimal SPM content and to be more strongly water soluble than the DDTs. The spread of HCH reflects this characteristic. We have observed that mean dissolved HCHs were significantly greater than its particulate form, as was to be anticipated. Even though the seasons with the highest concentrations of dissolved HCHs were summer and fall, their particulate content was at least ten times lower, causing a highly skewed distribution of HCHs in this period.

4.6.2 OCP sources in the SCS and ECS

As shown by the PCA findings (**Figure 4.5**), the dominance pattern varies greatly across seasons and stages. There was not a single species that loaded the main axis favourably. Rather, almost all of the HCB, HCH, and DDT species ascribed substantially to positive loadings of the first four axes explained more than 80% of the variance, suggesting numerous sources of OCPs in the SCS and ECS.

Agriculture could be the primary source of HCB if it is still used in this industry, despite the fact that it was banned in most countries, including China, in 2009[111]. It is currently released as a by-product or impurity in the production of chlorinated solvents (e.g., tetrachloroethylene, trichloroethylene, carbon tetrachloride), pesticides such as pentachloronitrobenzene, tetrachloroisophthalonitrile

(chlorothalonil), 4-amino-3,5,6-trichloropicolinic acid (picloram) [112-114]. It's also used in the manufacturing of atrazine, propazine, simazine, and mirex. Because it typically reacts with hydroxyl radicals, HCB has a high potential for long-range atmospheric transport, and its half-life period in the atmosphere is 7.7-14 years[115]. Due to its frequent interactions with hydroxyl radicals and its 7.7–14-year atmospheric half-life, HCB has a high capacity for long-range atmospheric movement [115]. Aside from binding to dirt and suspended particulates, HCB also travels through water. In comparison, particulate HCB contributed 32.42, 0.0, 25.16, and 10.04% of total particulate OCPs in the seasons of spring, summer, autumn, and winter, respectively. We observed that the total contribution of dissolved HCB was 23.63, 35.40, 7.47, and 19.24% of total dissolved OCPs. These significant spatiotemporal variations suggest that atmospheric deposition, along with other possible agricultural sources and commercial chemical manufacturing by products, may be a major source of HCB.

HCH contributions differed significantly between the stages and between the seasons. HCH contributed 37.68, 14.81, 52.28, and 18.16% of the total OCPs in the particulate phase, but accounted for 68.38, 42.61, 69.44, and 69.93% of the total OCPs in the dissolved phase in the spring, summer, autumn, and winter, respectively. Most HCH discharged into the atmosphere comes from lindane and technical HCH used in agriculture and medicine [116]. Technical HCH and lindane composition is usually represented by the isomers α, β, γ, δ-HCH. Technical HCH typically contains 60 to 70% α-HCH, 5 to 12% β -HCH, 10-15% γ -HCH, and 3 to 4 % δ-HCH and 3 to 4% other isomers, whereas lindane HCH contains more than 90% γ -HCH [117]. We discovered that α, β, γ, δ-HCH contributed differently depending on the season and

that their makeup significantly differed. Across seasons and stages, such suggestive compositional patterns have not been seen. However, the input pattern of the dissolved α-HCH (52.24%) >β -HCH (35.15%) > δ-HCH (7.0%) > γ -HCH (5.60%) hints at the use of technological HCH in the fall. The most common and stable variety of HCH is β-HCH, whereas α- and γ-HCH are readily converted to β-HCH [118, 119]. Moreover, photoisomerization can change γ-HCH into α-HCH [120]. Therefore, α/γ-HCH values can be used to locate possible HCH sources: <4, 4-7, and >7, respectively, suggest the use of lindane currently, technically HCH currently, and technically HCH historically[121]. It was observed that, in the spring, summer, autumn, and winter, respectively, dissolved α/γ-HCH ranged between 0.04-46.06, 0.11-33.49, 0.82-39.25, and 0.22-46.12. In the spring, summer, and winter, its levels stayed <4 in >75% of sampling sites; only in the autumn did it exhibit >7 in >75% of all sites. As a result, it suggested potential HCH sources for lindane's present or past expert HCH use. Parallel to this, particulate α/γ-HCH varied from 0.11 to 19.56 and displayed <4 in more than 80% of locations throughout all seasons. We computed the fraction β/(α+γ)-HCH to assess the degree of degradation. Over the course of the four seasons, this ratio varied between 0.11-29.75 and >1 in at least >31% of dissolved phase sites and 0.11-14.16 and >1 only in 40% of particulate phase sites, suggesting that HCH has not been severely deteriorated. This, along with the comparatively low α/γ-HCH ratio (<4) in all phases and seasons, suggested potential HCH sources from lindane use currently or from technical HCH use in the past.

In the spring, summer, autumn, and winter, respectively, DDTs made up 8.95, 21.99, 23.08, and 9.53% of all dissolved OCPs and 29.9, 83, 22.57, and 71.80% of all particle OCPs (**Table 4.4**) (**Table 4.5**). Individual DDT contributions varied greatly

between seasons and stages, and there was no clear season-wide supremacy of any one DDT species. We computed the ratio of DDE/DDD and (DDE+DDD)/DDT because DDTs frequently deteriorated into DDE and DDD, including p, p' and o, p' isomers. In the dissolved and particulate phases, DDE/DDD ranged between 0.11-18.35 and 0.11-35.23, and (DDE+DDD)/DDT between 0.23-527.32 and 0.11-238.46. The first ratio is <1 at least 25% sites and the second ratio is <1 at least 10% sites, which may suggest that DDTs in the majority of sites were not highly degraded. Additionally, the majority of the locations had modest ratios of, p'-DDT/p, p'-DDT (<0.25), suggesting that technical DDTs rather than dicofol may have been use.

Table 4.4 Source Analysis of Particulate OCPs

SPRING-HCH	α-HCH/	β-HCH	γ-HCH	δ-HCH	total	alpha/gamma	beta/(alpha+gamma)	% OF OCPS
Total sum	1.08934	0.76933	0.9709	0.10894	2.93851	NA	NA	37.6797894
%	37.0712	26.1811	33.0405	3.70716	100	NA	NA	
max						19.558899	1.749132389	
min						0.123323579	0.111336535	
% <4						95.40229885	NA	
% 4-7						2.298850575	NA	
% > 7						2.298850575	NA	
>1						NA	11.62790698	

(A)

SUMMER-HCH	α-HCH/	β-HCH	γ-HCH	δ-HCH	total	alpha/gamma	beta/(alpha+gamma)	% OF OCPS
Total sum	6.9242	6.01763	1.11577	9.19916	23.2568	NA	NA	14.8055793
%	29.7729	25.8747	4.79762	39.5548	100	NA	NA	
max						11.41689113	14.16125477	
min						0.119615312	0.127489369	
% <4						80.95238095	NA	
% 4-7						9.523809524	NA	
% > 7						9.523809524	NA	
>1						NA	37.87878788	

(B)

FALL-HCH	α-HCH/	β-HCH	γ-HCH	δ-HCH	total	alpha/gamma	beta/(alpha+gamma)	% OF OCPS
Total sum	1.60704	1.14165	1.52021	0.80041	5.06931	NA	NA	52.2759836
%	31.7014	22.5208	29.9885	15.7893	100	NA	NA	
max						16.292218 42	2.632554578	
min						0.16112894	0.112756916	
% <4						93.4210526 3	NA	
% 4-7						5.263157895	NA	
% > 7						1.315789474	NA	
>1						NA	12.5	

(C)

WINTER-HCH	α-HCH/	β-HCH	γ-HCH	δ-HCH	total	alpha/gamma	beta/(alpha+gamma)	% OF OCPS
-Total sum	0.18882	0.12186	0.19946	0.20329	0.71343	NA	NA	18.159207 9
%	26.4668	17.0812	27.9573	28.4947	100	NA	NA	
max						21.712368 76	7.612774878	
min						0.124176412	0.397658329	
% <4						89.189189 19	NA	
% 4-7						2.702702703	NA	
% > 7						8.108108108	NA	
>1						NA	40	

(D)

SPRING-DDTs	o,p'-DDE	p,p'-DDE	o,p'-DDD	p,p'-DDD	o,p'-DDT	p,p'-DDT	total	DDE/DDD	(DDE+DDD)/DDT	op'DDT/pp'DDT	% OF OCPS
Total sum	0.57976	0.49342	0.15266	0.88335	0.07803	0.144514902	2.331738659	NA	NA	NA	29.89931974
%	24.8638	21.161	6.54721	37.8839	3.34638	6.197731509	100	NA	NA	NA	
max								35.2314746	238.4581128	0.563465989	
min								0.11639885	0.319657487	0.121549198	
% <1								53.6585366	NA	NA	
% <1								NA	10.25641026	NA	
%< 0.25								NA	NA	61.53846154	

(E)

FALL-DDTs	o,p'-DDE	p,p'-DDE	o,p'-DDD	p,p'-DDD	o,p'-DDT	p,p'-DDT	total	DDE/DDD	(DDE+DDD)/DDT	op'DDT/pp'DDT	% OF OCPS
Total sum	0.60188	0.65254	0.03326	0.7538	0.14698	0	2.188461164	NA	NA	NA	22.56793397
%	27.5026	29.8175	1.51964	34.4443	6.71608	0	100	NA	NA	NA	
max								4.28464582	35.66151629	0	
min								0.25129673	1.474181486	0	
% <1								45.7627119	NA	NA	
% <1								NA	0	NA	
%< 0.25								NA	NA	0	

(F)

SUMMER-DDTs	o,p'-DDE	p,p'-DDE	o,p'-DDD	p,p'-DDD	o,p'-DDT	p,p'-DDT	total	DDE/DDD	(DDE+DDD)/DDT	op'DDT/pp'DDT	% OF OCPS
Total sum	31.5285	22.7126	23.1754	6.7196	29.3232	16.91718277	130.3765372	NA	NA	NA	82.99953354
%	24.1827	17.4208	17.7757	5.154	22.4912	12.97563807	100	NA-	NA	NA	
max								5.71861772	24.13969713	3.613593937	
min								0.32166825	0.119891842	0.135661343	
% <1								32.8571429	NA	NA	
% <1								NA	41.46341463	NA	
%< 0.25								NA	NA	10.41666667	

(G)

WINTER-DDTs	o,p'-DDE	p,p'-DDE	o,p'-DDD	p,p'-DDD	o,p'-DDT	p,p'-DDT	total	DDE/DDD	(DDE+DDD)/DDT	op'DDT/pp'DDT	% OF OCPS
Total sum	0.34228	0.73575	0.02364	0.31794	0.25385	1.147276547	2.820734627	NA	NA	NA	71.79712845
%	12.1343	26.0836	0.83818	11.2717	8.99928	40.67296457	100	NA	NA	NA	
max								5.575983832	61.47537829	2.185137847	
min								0.28872738	0.111613782	0.115664433	
% <1								38.461538 5	NA	NA	
% <1								NA	79.6875	NA	
%< 0.25								NA	NA	46.875	

(H)

HCB	Spring	Summer	Fall	Winter
Particulate (ng/L)	HCB	HCB	HCB	HCB
Total	2.52839	0	2.43944	0.39459
%	32.4209	0	25.1561	10.0437

Total OCPs			
Spring	Summer	Fall	Winter
7.798634481	157.0810481	9.69721538	3.928756885

(I)

Table 4.5 Source Analysis of Dissolved OCPs

Spring-HCH								
	α-HCH	β-HCH	γ-HCH	δ-HCH	total	alpha/gamma	beta/(alpha+gamma)	% of total OCPs
Total sum	27.12941011	77.67002297	27.28571672	29.04413304	161.1292828	NA	NA	68.37968645
%	16.83707471	48.20362751	16.93408184	18.02539148	100	NA	NA	
max						46.05558328	29.74791639	
min						0.040988362	0.118133393	
% <4						82.66666667	NA	
% 4-7						6.666666667	NA	
% > 7						10.66666667	NA	
%>1							58.13953488	

(A)

Summer-HCH								
	α-HCH/	β-HCH	γ-HCH	δ-HCH	total	alpha/gamma	beta/(alpha+gamma)	% of total OCPs
Total sum	117.7751893	149.6394129	63.01162026	7.112197494	337.53842	NA	NA	42.61010238
%	34.89238242	44.33255977	18.66798571	2.107078037	100	NA	NA	
max						33.48652377	19.91815111	
min						0.111478245	0.118812476	
% <4						75.67567568	NA	
% 4-7						8.108108108	NA	
% > 7						16.21621622	NA	
>1							31.91489362	

(B)

Fall-HCH								
	α-HCH/	β-HCH	γ-HCH	δ-HCH	total	alpha/gamma	beta/(alpha+gamma)	% of total OCPs
Total HCH	132.8747757	89.41735328	14.25179699	17.81314293	254.3570689	NA	NA	69.44363118
%	52.23946008	35.15425883	5.603066315	7.003202557	100	NA	NA	
max						39.25235043	2.883127325	
min						0.821297148	0.336631624	
% <4						8.974358974	NA	
% 4-7						15.38461538	NA	
% > 7						75.64102564	NA	
>1						NA	11.53846154	

(C)

Winter-HCH								
	α-HCH/	β-HCH	γ-HCH	δ-HCH	total	alpha/gamma	beta/(alpha+gamma)	% of total OCPs
Total HCH	42.42624593	111.3582222	15.92887135	20.92348812	190.6368276	NA	NA	69.9346089
%	22.25501368	58.41381214	8.355612006	10.97557666	100.0000145	NA	NA	
max						46.12142439	19.65736553	
min						0.219457648	0.253648437	
% <4						79.74683544	NA	
% 4-7						7.594936709	NA	
% > 7						12.65822785	NA	
>1						NA	82.27848101	

(D)

Spring-DDTs											
	o,p'-DDE	p,p'-DDE	o,p'-DDD	p,p'-DDD	o,p'-DDT	p,p'-DDT	total	DDE/DDD	(DDE+DDD)/DDT	op'DDT/pp'DDT	% of total OCPs
Total DDTs	3.795673258	3.537338975	4.08830575	2.812640836	2.01194461	4.846446979	21.09235041	NA	NA	NA	8.951124724
%	17.99549722	16.77072007	19.38288408	13.33488604	9.538740872	22.97727365	100	NA	NA	NA	
max								8.725795155	28.52187967	3.222224448	
min								0.112286824	0.216475613	0.125487843	
% <1								52.38095238	NA	NA	
% <1								NA	44.82758621	NA	
%< 0.25								NA	NA	50	

(E)

Summer-DDTs											
	o,p'-DDE	p,p'-DDE	o,p'-DDD	p,p'-DDD	o,p'-DDT	p,p'-DDT	total	DDE/DDD	(DDE+DDD)/DDT	op'DDT/pp'DDT	% of total OCPs
Total DDTs	76.676395 62	13.070635 16	72.589937 32	5.6641608 36	0.5090182 87	5.7018327 16	174.21197 99	NA	NA	NA	21.99213441
%	44.013268 67	7.5027180 47	41.667587 38	3.25130349	0.2921832 52	3.2729276 49	100	NA	NA	NA	
max								8.949238622	4.316996255	0.221758866	
min								0.113216295	0.23745414	0.124425467	
% <1								60	NA	NA	
% <1								NA	42.85714286	NA	
%< 0.25								NA	NA	100	

(F)

Fall-DDTs											
	o,p'-DDE	p,p'-DDE	o,p'-DDD	p,p'-DDD	o,p'-DDT	p,p'-DDT	total	DDE/DDD	(DDE+DDD)/DDT	op'DDT/pp'DDT	% of total OCPs
Total DDTs	10.00117955	18.71062401	13.01723525	40.91375714	0.813397519	1.090306949	84.54650041	NA	NA	NA	23.0825745
%	11.82920582	22.13057183	15.3965394	48.39201734	0.962071186	1.289594417	100	NA	NA	NA	
max								18.3586759	527.3159993	4.463335166	
min								0.132227532	1.722148293	0.155452978	
% <1								37.70491803	NA	NA	
% <1								NA	0	NA	
%< 0.25								NA	NA	5.882352941	

(G)

Winter-DDTs											
	o,p'-DDE	p,p'-DDE	o,p'-DDD	p,p'-DDD	o,p'-DDT	p,p'-DDT	total	DDE/DDD	(DDE+DDD)/DDT	op'DDT/pp'DDT	% of total OCPs
Total DDTs	17.481876311	3.678046022	0.319640384	1.195479809	0	3.31155394	25.98659646	NA	NA	NA	9.533113212
%	67.272657009	14.153625441	1.230020024	4.600370224	0	12.74331363	100	NA	NA	NA	
max								7.235811653	4.118678327	0	
min								0.46844143	0.552324923	0	
% <1								25	NA	NA	
% <1								NA	50	NA	
%< 0.25								NA	NA	0	

(H)

HCB				
Dissolved (ng/L)	Spring	summer	fall	Winter
sites	HCB	HCB	HCB	HCB
Total	55.68259204	280.4054531	27.37489942	52.4389418
%	23.63045524	35.39776321	7.473794327	19.23708515

(I)

Total OCPs			
Spring	summer	fall	winter
235.6391075	792.1558531	366.2784688	272.5929702

(J)

4.6.3 Maps of ecological risk

MRM successfully assessed the danger presented to an ecosystem by a chemical mixture rather than a single substance. In the current research, a two-tier MRM was used to evaluate the ecological risk of all 11 OCPs in the SCS and ECS. Risk profiles in dissolved and particulate stages were created for each of the OCP groups in four seasons based on its production. A similar model was used to evaluate the risk of 15 OCPs in the surface water of the Qingshitan reservoir in Southwest China and discovered a high potential risk to the aquatic ecosystem [91]. However, seasonal and phase-partitioning impacts were not taken into account in this evaluation.

The current research specifically examined the impacts of season and phase partitioning by creating separate maps (**Figure 4.7**) for each season and phase. Such analysis revealed that dissolved HCB and DDTs presented low-to-medium levels of possible risk throughout the seasons, whereas only particulate DDTs posed a comparable risk during the summer season. Only a few locations indicated high-risk potential for HCB and DDTs during the summer. Notably, HCH presented a minimal degree of risk in all seasons in a few locations. Similarly, a comprehensive study of

OCPs in South American settings found that the majority of locations are risk-free for biota [122]; however, some OCPs may be of concern for possible harm to ecosystem structure and functioning.

.

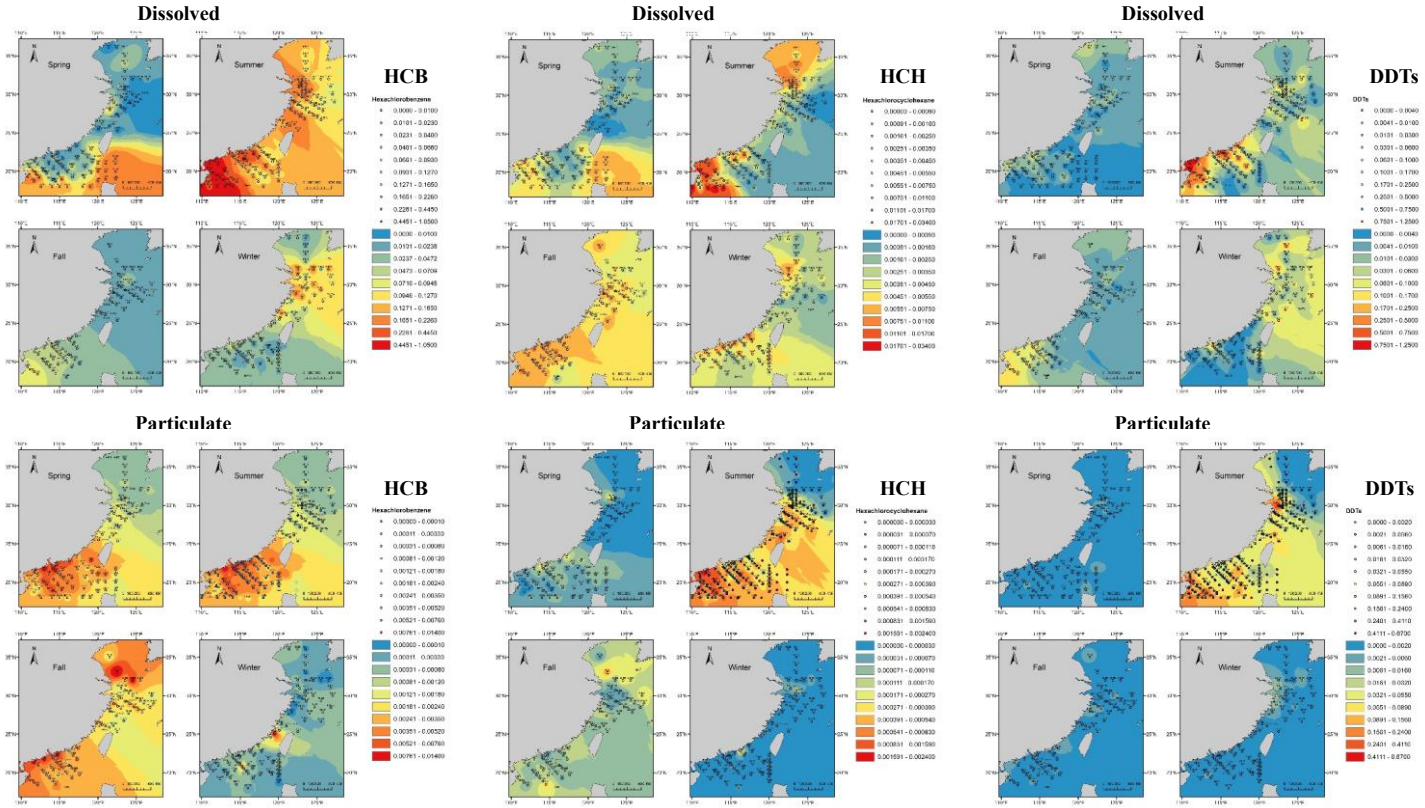


Figure 4.7 Spatiotemporal ecological risk maps

Spatiotemporal ecological risk maps: It showed that spatiotemporal variation of the risks posed by three primary OCP classes, HCB, HCH, and DDTs in SCS and ECS. Dissolved HCB and DDTs posed low-to medium levels of potential risk widely spread across the seasons; whereas only particulate DDTs displayed similar risk in summer season. Very few spots showed high-risk potential for HCB and DDTs in summer only. Notably, HCH posed low-level of risk in all the seasons in few sites (Low risk: 0.01-0.1; Medium risk: 0.1-1.0; High risk: >1.0).

4.7 Summary

5. In this chapter I have conducted a spatial temporal data analysis of OCPs concentration of South China Sea (SCS) and East China sea (ECS). This research concentrated on the ecological risk evaluation of OCPs (organochlorine pesticides) seasonal and phase-petitioning effects in SCS and ECS, which are important source-sink zones.

6. The findings revealed significant temporal heterogeneity in OCP distribution, with significant differences between the dissolved and particulate phases. PCA analysis indicated numerous OCP sources, including present and past use of HCH and DDT, atmospheric transport, and HCB deposition from terrestrial surfaces.

7. The spatiotemporal ecological risk analyses revealed no high-risk zones, but did identify one or two high-risk regions for HCB and DDTs. Dissolved and particulate OCPs presented low-to-medium amounts of risk, with summer posing a slightly greater risk.

8. Overall, the research emphasises the significance of taking spatiotemporal variation into account when assessing ecological risk.

“Time series analysis is like a microscope through which we can see the hidden properties of data”. … George Box

Chapter 5

5. Time series data modelling and analysis

5.1 Introduction

Malaria, a vector-borne and one of the most lethal Plasmodium species-caused illness, kills one person every minute worldwide and has a devastating effect on people's health and income [123]; [124];[125]. It is estimated that 4 billion people are at risk in 87 countries, with 229 million cases, killing nearly 409,000 people, mostly children under the age of five, in Sub-Saharan Africa in 2019 [126]; [127]),and Southeast Asia has a large percentage of young teenagers at danger [128];[129];[130]. The West and African regions have recorded a huge load of malaria morbidity, accounting for 95% of all malaria cases worldwide, particularly in one of the world's lowest geographic regions, followed by Southeast Asia with 5 million cases and 9,000 fatalities [131]; [132].

Wang, C., **Thakuri, B**.*, Roy, A. K., Mondal, N., Qi, Y., Chakraborty, A. (2023). Changes in the associations between malaria incidence and climatic factors across malaria endemic countries in Africa and Asia-Pacific region. *Journal of Environmental Management*, *331*, 117264

Significant differences in malaria morbidity are reported within and between the African and Asia-Pacific regions, with Nigeria (31.9%), the Democratic Republic of the Congo (13.2%), the United Republic of Tanzania (4.1%), and Mozambique (3.8%) accounting for the highest prevalence of malaria deaths in the African region, and India alone accounting for 82.5%, Indonesia (15.6%), and Myanmar (1.6%) in Southeast Asia ([133]; [134].

This variety in malaria prevalence and mortality has stayed crucial in the WHO's global malaria control and eradication efforts, which resulted in the audacious Global Malaria Program (GMP) with the goal of eradicating 90% of the world malaria load by 2030 ([135]; [136]. The duties of the GMP are guided by the global technical strategy for malaria 2016-2030, which was approved by the 66th World Health Assembly in May 2015 and was recently revised in 2021 to address the shifting malaria environment. The most current WHO World Malaria Report 2021 included projections of the effect of the COVID-19 pandemic on critical malaria services. According to the study, malaria deaths rose by 12% in 2020 compared to 2019, to an estimated 627,000 worldwide. During the COVID-19 pandemic, malaria service delays were responsible for an estimated 47,000 (68%) of the extra 69,000 fatalities.

Malaria is extremely susceptible to climatic variables such as temperature, precipitation, and humidity[137-139], which influence the sporogony cycle of Plasmodium species directly and indirectly through mosquito-human interactions. Several studies have verified that temperature changes have a significant impact on malaria spread[140-142]. As a result, it has become a major worry as global temperatures have risen considerably over the last 100 years. A simple modelling research concludes that rising temperatures will increase malaria transmission and

broaden its regional dispersal [143]. Furthermore, the rise in malaria prevalence is favourably linked to the quantity of rainfall, owing to an increase in mosquito breeding locations, and thus the volume grows with rainfall[144, 145]. Multiple studies have also found that the illness has resurfaced as a result of global warming, climate change, and human activities in an area where it has previously been effectively managed or eradicated[146-148]. Furthermore, several recent studies have found a close relationship between global warming and climate change and malaria incidence [149-151], demonstrating that rising temperatures can have opposing effects on malaria dynamics in both highland and lowland regions[152-155]. Improvements in socioeconomic conditions, improved irrigation and living conditions, contemporary agricultural techniques, house screening, technical development, and access to better healthcare facilities, on the other hand, played critical roles in containing the malaria prevalent in impacted[156-158]. Because of large intra- and inter-regional differences in malaria incidence and a lack of empirical evidence, the link between malaria incidence and climatic factors across countries has remained controversial, as socioeconomic development has frequently been found to outweigh the climatic effects at the country level [159] raising questions about the extent of the relationship in the presence of various other confounding factors. The current research aimed to assess the relationships between malaria incidence and climatic factors across nations and measure the impacts of intrinsic large variations in climatic factors.

5.2 Methods

5.2.1 Data collection

Malaria incidence data were gathered from the World Health Organization's Global Health Observatory Data Repository for the years 2000-2020 (https://ourworldindata.

org/malaria). The amount of new instances of malaria per 1000 people at risk is referred to as the incidence. The two most malaria-endemic areas, Africa and the Asia-Pacific zone, were chosen. To address our specific research question of how the type and breadth of the correlations between malaria prevalence and climatic factors (temperature and precipitation) vary across nations The current research took into account malaria-affected 42 African countries and 20 Asia-Pacific countries as stated by WHO. Only malaria-endemic nations with comparatively modest variation in mean annual temperature over the last two decades were considered. Particularly, all of the nations studied showed a narrow temperature range of 12-30°C with a standard deviation (SD) of < 0.5. Few outliers that had been considered are "China" and "North Korea", which revealed the average yearly temperature of 7.56 ± 0.27℃, and 7.02 ± 0.48℃ respectively, due to their broad regional influence and high cases of malaria frequency in the past. This exception, in particular, of China, has been clearly represented in our study, which shows a totally distinct magnitude of the link between peers. Few malaria-affected nations from either area were excluded due to very high variability in mean-annual temperature (i.e., SD > 0.5) or because they were not mentioned in the WHO malaria report. The dataset's African malaria endemic region, Few malaria-affected nations from either area were excluded due to very high variability in mean-annual temperature (i.e., SD > 0.5) or because they were not mentioned in the WHO malaria report.

The African malaria endemic region of the dataset includes Angola (AGO), Burundi (BDI), Benin (BEN), Burkina Faso (BFA), Botswana (BWA), Central African Republic (CAF), Cote d'Ivoire (CIV), Cameroon (CMR), Cape Verde (CPV), Chad (TCD), Republic of the Congo (COG), Democratic Republic of Congo (COD),

Comoros (COM), Ethiopia (ETH), Equatorial Guinea (GNQ), Gabon (GAB), Ghana (GHA), Guinea (GIN), The Gambia (GMB), Guinea-Bissau (GNB), Kenya (KEN), Liberia (LBR), Madagascar (MDG), Mali (MLI), Mozambique (MOZ), Mauritania (MRT), Malawi (MWI), Namibia (NAM), Niger (NER), Nigeria (NGA), Rwanda (RWA), Sudan (SDN), Senegal (SEN), Sierra Leone (SLE), Somalia (SOM), South Sudan (SSD), Sao Tome and Principe (STP), Togo (TGO), United Republic of Tanzania (TZA), Uganda (UGA), South Africa (ZAF), and Zambia (ZMB). Among these nations, NGA has the most people (206,139,587) and the largest territory (9,23,768 km^2), while STP has the smallest area (964 km^2) and the fewest people (2,23,107). Malaria-impacted countries of Asia-Pacific region are Afghanistan (AFG), Bangladesh (BGD), Bhutan (BTN), China (CHN), Indonesia (IDN), India (IND), South Korea (KOR), Sri Lanka (LKA), Myanmar (MMR), Malaysia (MYS), Nepal (NPL), Pakistan (PAK), Philippines (PHL), Papua New Guinea (PNG), North Korea (PRK), Solomon Islands (SLB), Thailand (THA), Timor-Leste (TLS), Vietnam (VNM), and Vanuatu (VUT). There is a significant difference in area and population number, while China and India together account for about 36% of total global population and 67% of Asia population, respectively.

According to the WHO World Malaria Report 2021, three distinct techniques were used for country-by-country estimation of malaria cases from 2000 to 2020, which properly accounted for the unavoidable uncertainties surrounding the number of cases. Method 1 was used for nations and regions outside of the WHO African zone with minimal malaria transmission. It contains AFG, BGD, BWA, ETH, GNB, IND, IDN, MDG, MRT, MMR, NAM, NPL, PAK, PNG, PHL, RWA, SEN, SLB, TLS, VUT, and VNM, with estimates adjusted for completeness of reporting, the

probability of parasite positive cases, and the amount of health care use. Method 2 was used because high transmission countries in the WHO African and Eastern Mediterranean regions lacked quality surveillance data; these countries include AGO, BEN, BFA, BDI, CMR, CAF, TCD, COG, COD, GNQ, GAB, GHA, GIN, GNB, KEN, LBR, MWR, MLI, MOZ, NER, NGA, SLE, SOM, SSD, TGO, UGA, TZA, and ZMB. Method 2 calculates the number of malaria cases using parasite prevalence data from community questionnaires. It employed a spatiotemporal Bayesian geostatistical model, as well as environmental and socioeconomic variables, as well as intervention data such as antimalarial medicines, residual sprinkling, and insecticide-treated mosquito nets. Method 3 was used for nations in the protection stage of reintroduction, and it used local cases recorded by the National Malaria Program. It consists of the following countries: BTN, CPV, CHN, COM, PRK, MYS, KOR, STP, ZAF, LKA, and THA.

To link malaria incidence to climatic variables, statistics on yearly min, max, mean temperature (°C/year/country), and precipitation (mm/year/country) were obtained from the World Bank group's Climate Change Knowledge Portal (CCKP) (https://climateknowledgeportal.worldbank.org). For the same time span of 2000-2020, it used the data source CRU TS v4.05 (Climatic Research Centre Gridded Time Series). Except for Antarctica, this is the most commonly used observational climate record produced on a 0.5° latitude by 0.5° longitude grid. CRU TS climate data were generated by interpolating monthly temperature anomalies from large networks of weather station records.

5.2.2 Generalized linear model and mixed effects model

A linear connection between malaria frequency, weather, and precipitation is investigated using the Generalized linear model (GLM). The fundamental connection between the response (malaria incidence) and predictors (temperature and precipitation) may not be linear, as demonstrated by the use of a link function that links the response variable to a linear model. GLM was initially run independently for the two areas, without categorising temperature and precipitation data. The GLM-G (generalised linear model with group data) was then performed for each nation as a category.

Before providing inputs to the models, appropriate scaling of response and predictor variables is selected to deal with inherent heterogeneities in the dataset effectively and build up a comparable level. Because malaria incidences vary greatly across nations and between areas, standard log transforms of malaria incidence are used to decrease all values by <10 and the absolute deviation by 1.15. Malaria incidence, yearly *minimum temperature, and precipitation have shown a strong correlation between years, which may influence the statistical significance of model fitting. To prevent these possible connections, the annual lowest temperature and precipitation are adjusted by dividing the annual values by the country-wide maximum and then subtracting the mean values. The models are fitted using these scaled predictor factors to determine intercepts, slopes, standard errors (SEs), and p-values (distribution: Gamma). Because a recent study found that minimum temperatures (usually measured around sunrise) increase quicker over time than maximum (daytime) temperatures [160] we used annual minimum temperature rather than mean temperature to successfully fit the data into the models:

GLM: $\frac{1}{\text{Log}(10^6 \times \text{incidence})} = \alpha_0 + \alpha_1$ **(scaled min. temperature)** $+\alpha_2$ **(scaled precipitation)** *5.1*

GLM-G: $\frac{1}{\text{Log}(10^6 \times \text{incidence})} = \alpha_0 + \alpha_1$ **(country: scaled min. temperature)** $+\alpha_2$ **(country:scaled precipitation)** *5.2*

The Generalized Linear Mixed-Effects model (GLME) is a GLM modification with category or group dataset inputs. It distinguishes between fixed and random impacts. The fixed-effects term typically refers to the traditional linear regression component of the model, whereas the random-effects term is linked with experimental units chosen at random from a population, allowing for differences between groups that influence model performance. In order to suit the GLME, the *fitglme* inbuilt function in MATLAB R2021a is used:

Log $(10^6 \times$ **incidence)** $= \beta_0 + \beta_1$ **(scaled min. temperature)** $+\beta_2$ **(scaled precipitation)** $+ (\alpha_0 + \alpha_1$ **scaled min. temperature** $+ \alpha_2$ **scaled precipitation | country).** *5.3*

5.2.3 Agglomerative clustering

This is the most common clustering method, in which numerous items are grouped into clusters based on their similarity. It employs a bottom-up strategy. Initially, every object is regarded as a singleton collection or a cluster. At each recursive stage, the two most similar clusters are grouped to create a new larger cluster. This process is repeated until all of the data points form a single big cluster. For the country-wise GLM-G fixed values of temperature and precipitation coefficients, we used the agglomerative clustering algorithm in Scikit-learn 1.2.0 with the “Euclidian” distance

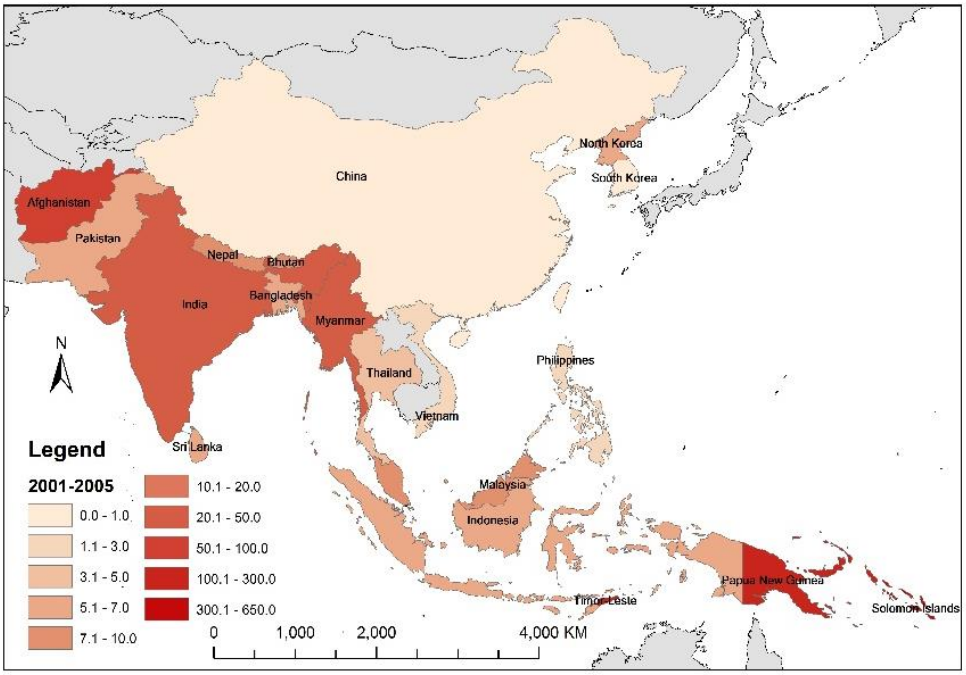
measure. To reduce the variance of the groups being combined, a linkage function called “ward” was used.

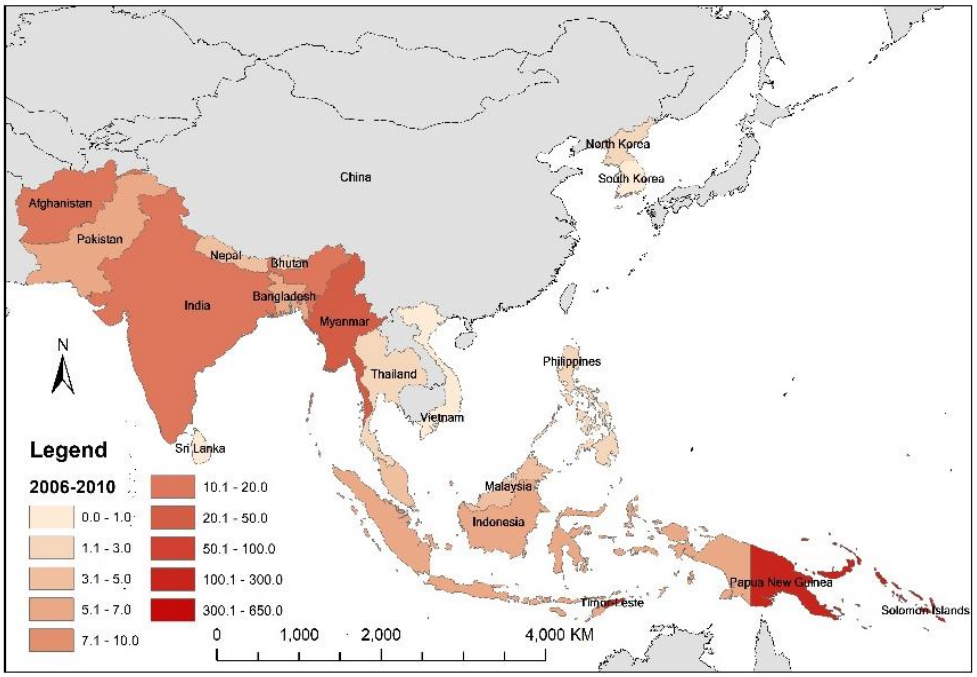
5.3 Results

5.3.1 Malaria prevalence and climatic variables vary intra- and inter-regionally.

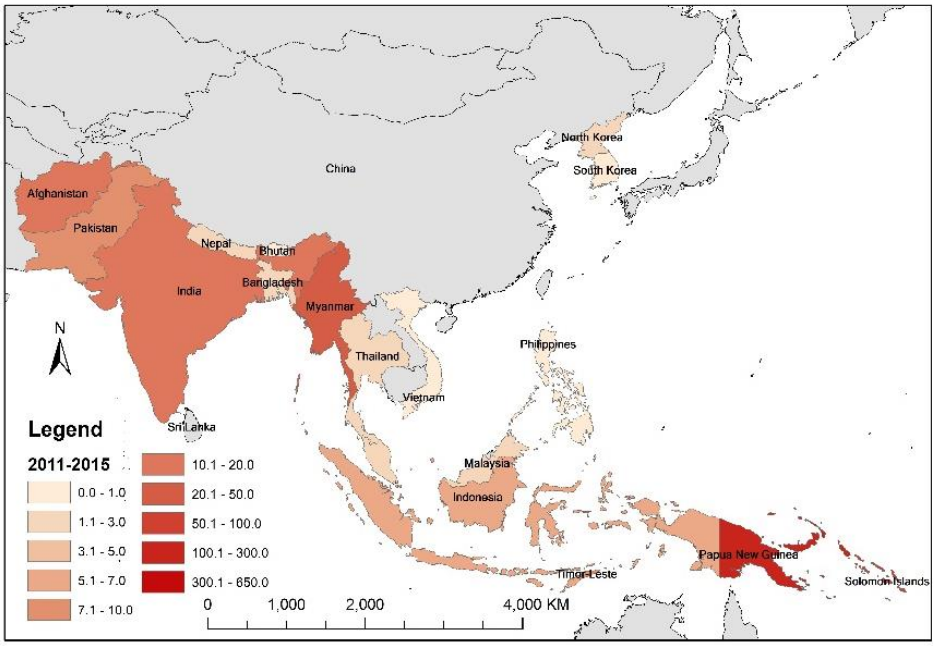
Malaria incidence per 1000 people varied greatly across nations and between Africa and the Asia-Pacific area (**Figure 5.1**). RWA had the greatest incidence of 724.57 (in 2017) with a yearly mean of 222.44 and the largest deviation of 161.82 in Africa between 2000 and 2020. In the last two decades, BFA had the greatest mean incidence of 502.54± 85.97, while 27 nations out of 42 had more than 200 mean incidence, including MOZ, UGA, and BDI. From 2006 to 2020, STP was the only nation that continuously demonstrated a <100 incidence rate, with an annual mean of 91.91± 114.27. Furthermore, only three African nations, BWA, ZAF, and CPV, had a mean incidence of <5. Across all African nations, substantial yearly deviations in the range 0.67-161.82 were found between 2000 and 2020, suggesting significant inter-country variation in malaria incidence. Except for SLB (310.30± 238.27), PNG (183.31 4± 3.39), TLS (83.56 ± 78.08), VUT (76.30 ± 68.16), MMR (33.33 ± 21.50), AFG (29.62 ± 24.17), and IND (14.10 ± 5.92), Asia-Pacific nations had at least ten times lower mean annual incidence over the same era. SLB constantly had an annual malaria incidence of more than 50 cases, with the maximum number of 744.16 in 2004, whereas PNG consistently had more than 100 cases of malaria incidence. While TLS, VUT, MMR, and AFG had yearly incidences greater than 25 from 2000 to 2005, IND had reliably had incidences less than 25 for the previous two decades. IND observed a

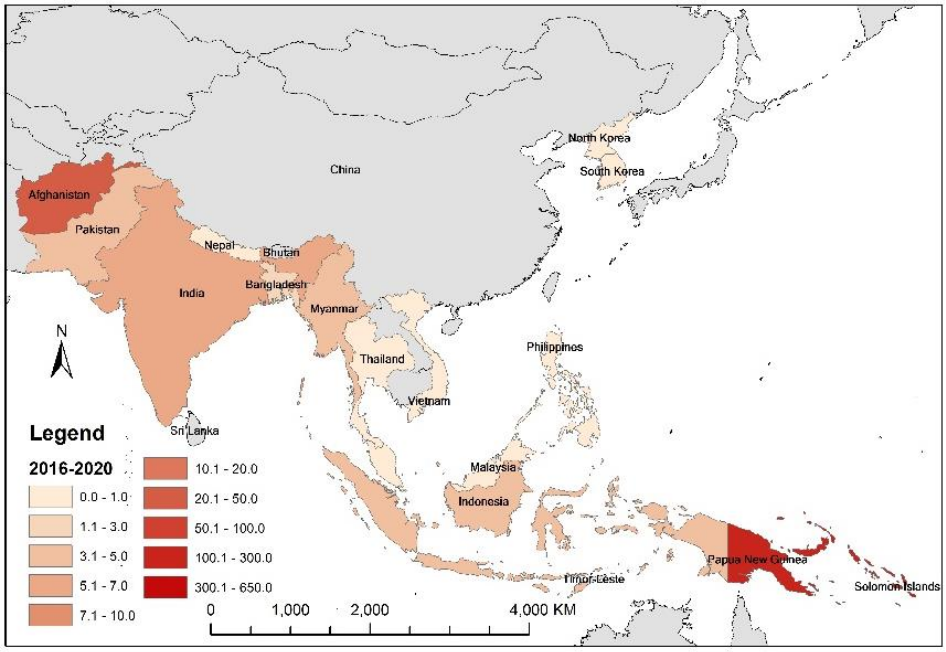
strict downward pattern from 19.91 in 2000 to 3.33 in 2020. LKA and AFG both had large yearly deviations (i.e., >10) over the last two decades, though LKA had a much lower mean incidence number of 5.97. The box plots depicted the spread of malaria incidence over the last two decades in Africa and Asia-Pacific nations (**Figure 5.2**). There is rare overlapping of confidence intervals, suggesting inherent malaria heterogeneity within and between areas. To assess decadal changes in each nation, the 5-year average malaria incidence over the last 20 years was computed. Malaria prevalence revealed significant intra- and inter-regional variations, with almost purely decadal declining patterns in Asia-Pacific and a kind of mixed trend in Africa (**Figure 5.1**).


China
North Korea
South Korea
Afghanistan
Pakistan
Nepal
Bhutan
Bangladesh
India
Myanmar
Thailand
Philippines
Vietnam
Sri Lanka
Malaysia
Indonesia
Papua New Guinea
Timor-Leste
Solomon Islands
N
Legend
2001-2005
0.0 - 1.0
1.1 - 3.0
3.1 - 5.0
5.1 - 7.0
7.1 - 10.0
10.1 - 20.0
20.1 - 50.0
50.1 - 100.0
100.1 - 300.0
300.1 - 650.0
0
1,000
2,000
4,000 KM

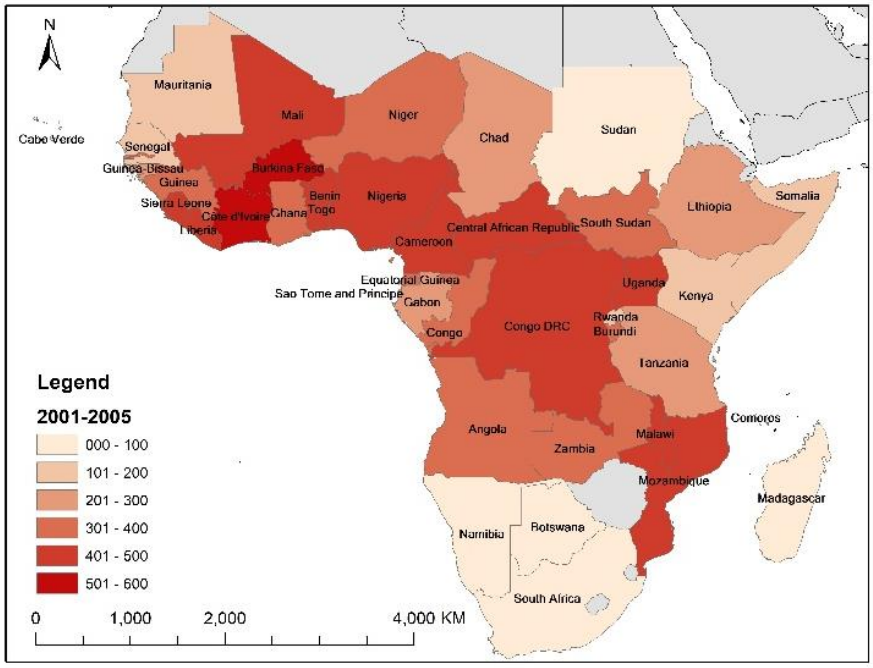

China
North Korea
South Korea
Afghanistan
Pakistan
Nepal
Bhutan
Bangladesh
India
Myanmar
Thailand
Philippines
Vietnam
Sri Lanka
Malaysia
Indonesia
Papua New Guinea
Timor-Leste
Solomon Islands
N
Legend
2006-2010
0.0 - 1.0
1.1 - 3.0
3.1 - 5.0
5.1 - 7.0
7.1 - 10.0
10.1 - 20.0
20.1 - 50.0
50.1 - 100.0
100.1 - 300.0
300.1 - 650.0
0
1,000
2,000
4,000 KM

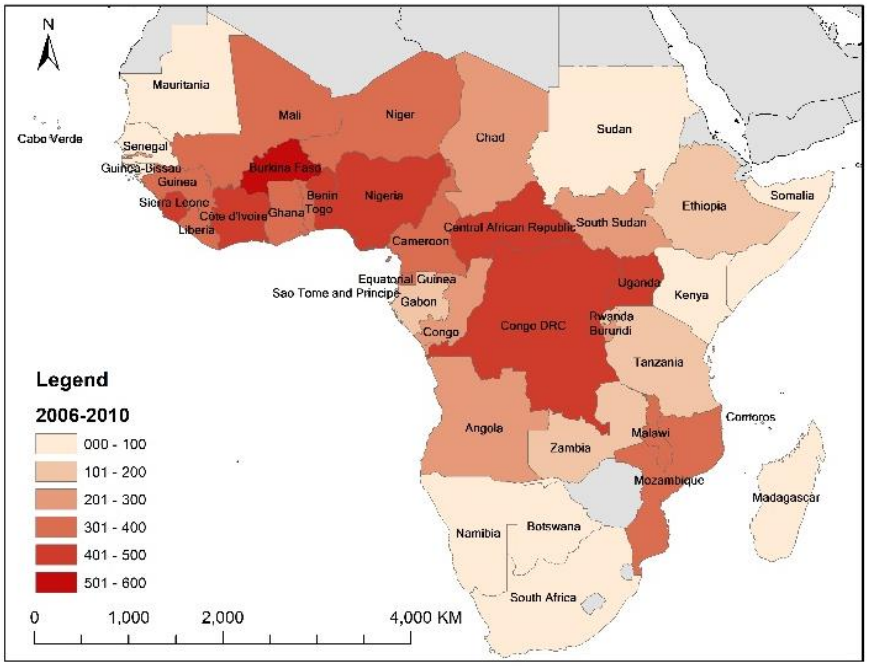
(A)


China
North Korea
South Korea
Afghanistan
Pakistan
Nepal
Bhutan
India
Bangladesh
Myanmar
Thailand
Vietnam
Philippines
Sri Lanka
Malaysia
Indonesia
Papua New Guinea
Timor-Leste
Solomon Islands
N
Legend
2011-2015
0.0 - 1.0
1.1 - 3.0
3.1 - 5.0
5.1 - 7.0
7.1 - 10.0
10.1 - 20.0
20.1 - 50.0
50.1 - 100.0
100.1 - 300.0
300.1 - 650.0
0
1,000
2,000
4,000 KM

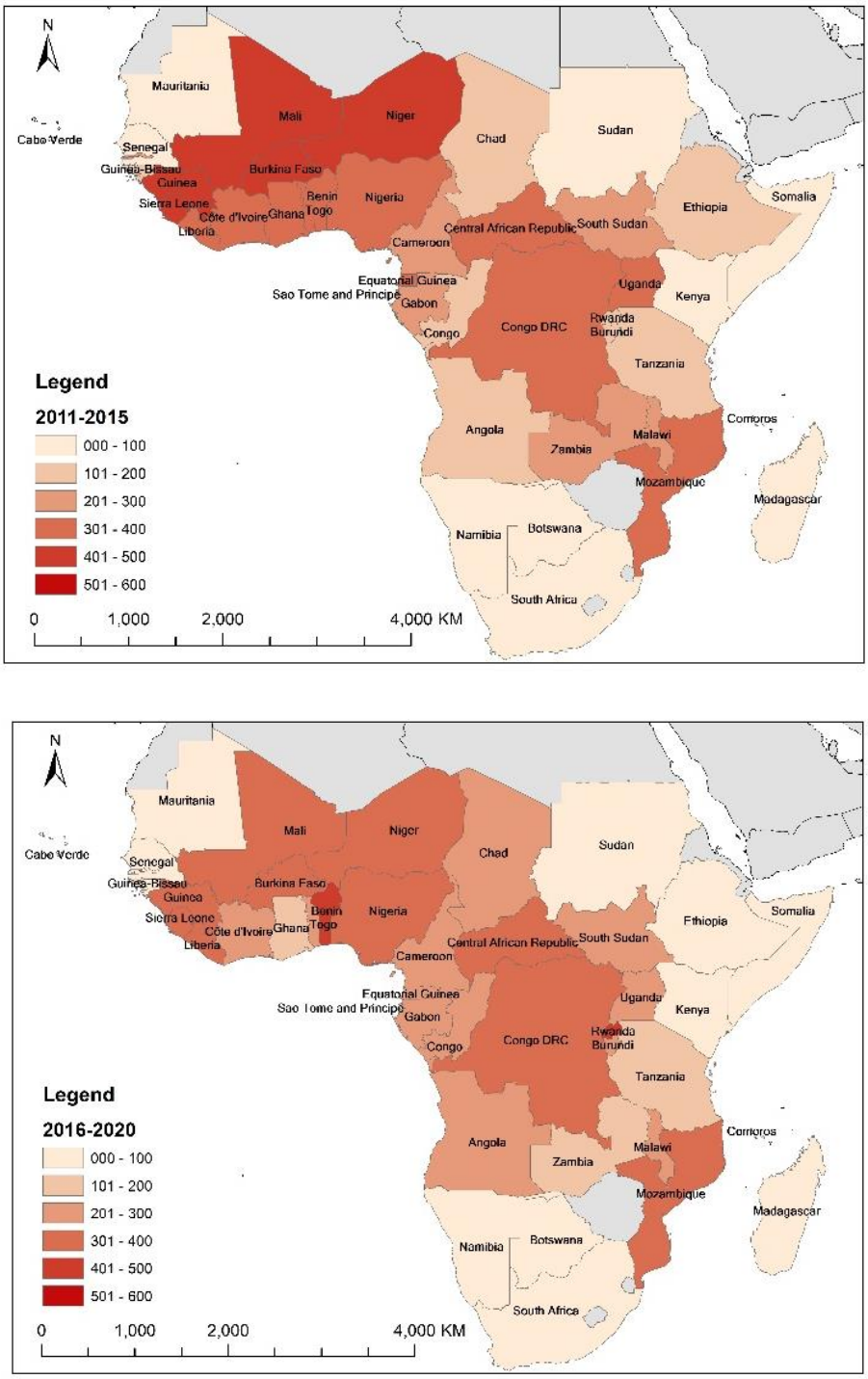

China
North Korea
South Korea
Afghanistan
Pakistan
Nepal
Bhutan
India
Bangladesh
Myanmar
Thailand
Vietnam
Philippines
Sri Lanka
Malaysia
Indonesia
Papua New Guinea
Timor-Leste
Solomon Islands
N
Legend
2016-2020
0.0 - 1.0
1.1 - 3.0
3.1 - 5.0
5.1 - 7.0
7.1 - 10.0
10.1 - 20.0
20.1 - 50.0
50.1 - 100.0
100.1 - 300.0
300.1 - 650.0
0
1,000
2,000
4,000 KM

(B)


N
Mauritania
Mali
Niger
Chad
Sudan
Cabo Verde
Senegal
Guinea-Bissau
Guinea
Burkina Faso
Benin
Nigeria
Sierra Leone
Côte d'Ivoire
Ghana
Togo
Liberia
Ethiopia
Somalia
Central African Republic
South Sudan
Cameroon
Equatorial Guinea
Sao Tome and Principe
Gabon
Uganda
Kenya
Congo
Congo DRC
Rwanda
Burundi
Tanzania
Comoros
Angola
Malawi
Zambia
Mozambique
Madagascar
Namibia
Botswana
South Africa
Legend
2001-2005
000 - 100
101 - 200
201 - 300
301 - 400
401 - 500
501 - 600
0
1,000
2,000
4,000 KM


N
Mauritania
Mali
Niger
Chad
Sudan
Cabo Verde
Senegal
Guinea-Bissau
Guinea
Burkina Faso
Benin
Nigeria
Sierra Leone
Côte d'Ivoire
Ghana
Togo
Liberia
Ethiopia
Somalia
Central African Republic
South Sudan
Cameroon
Equatorial Guinea
Sao Tome and Principe
Gabon
Uganda
Kenya
Congo
Congo DRC
Rwanda
Burundi
Tanzania
Comoros
Angola
Malawi
Zambia
Mozambique
Madagascar
Namibia
Botswana
South Africa
Legend
2006-2010
000 - 100
101 - 200
201 - 300
301 - 400
401 - 500
501 - 600
0
1,000
2,000
4,000 KM

(C)



(D)

Figure 5.1 Two most Malaria -affected global areas, Asia Pacific and Africa

Figure 5.1 depicts the two regions of the world most impacted by malaria: Asia-Pacific (Figure 5.1 A, B) and Africa (Figure 5.1 C, D). Twenty countries in Asia and the Pacific and 42 countries in Africa are both affected by malaria. Over the past two decades (2000-2020), all of the malaria-affected countries in both areas showed very minor variations in mean annual temperature, with a range of 12-300 C and a standard deviation (SD) of 0.5. Significant intra- and inter-regional variation in malaria incidence during the past 20 years was found, with almost exclusively decadal lowering patterns in Asia-Pacific and a somewhat mixed trend in Africa.

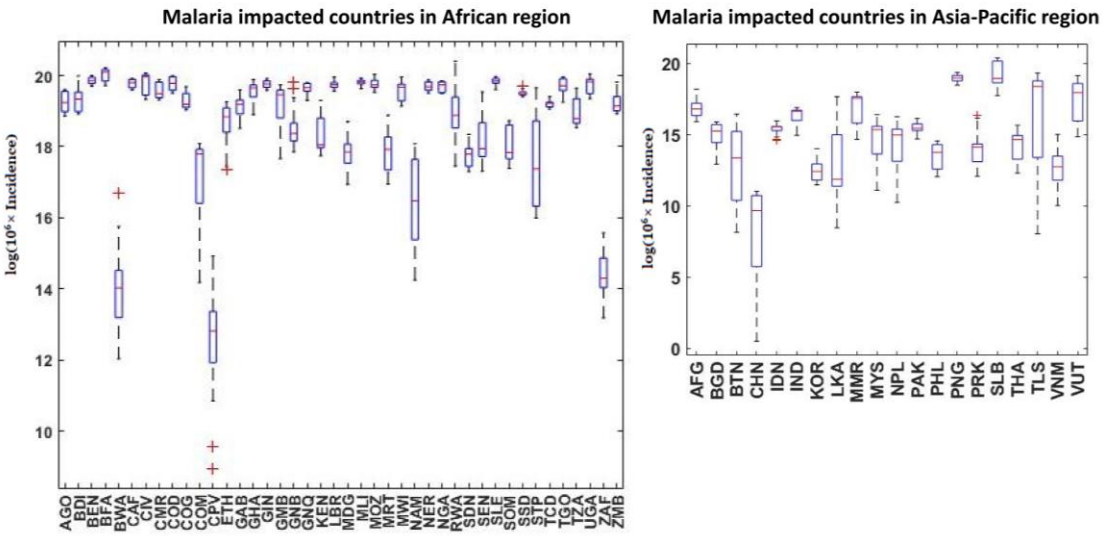


Figure 5.2 Distribution of Malaria incidence

The distribution of malaria incidence during the past 20 years, with 42 and 20 countries, respectively, in the African and Asia-Pacific regions: It showed a rare overlap of confidence intervals, pointing to inherent regional and intra-regional malaria heterogeneity.

Several climatic variables, including temperature and precipitation, are frequently linked to malaria incidence. During the period 2000-2020, yearly mean air temperatures in Africa fluctuated in a very narrow range of 18°C to 30°C, with a deviation of <0.4°C. Nations with mean yearly temperatures greater than 28°C

included BEN, BFA, GMB, GNB, MLI, MRT, NER, SEN, SSD, and SDN. Annual precipitation, on the other hand, varied greatly across all African nations during the same time period, with the inter-annual deviation of mean precipitation ranging from 14 -319 millimetres. COG, COD, GNQ, GAB, STP, CMR, COM, GNB, GIN, LBR, MDG, and SLE had recorded annual precipitation of more than 1500 mm almost every year during that time frame. Both annual air temperature and precipitation revealed comparatively significant variance in Asia-Pacific among countries with high annual deviations, suggesting greater inter-country climatic heterogeneity. Annual mean temperatures ranged from 7.51-27.36°C over a 20-year span, with deviations ranging from 0.04-0.48°C. During that time, the annual mean weather in BGD, IND, LKA, IDN, MYS, THA, PHL, VNM, and SLB was regularly above 25°C. North Korea and China had the lowest mean yearly weather over the last 20 years (~7°C). In comparison to African nations, Asia-Pacific countries had a significant variance in yearly precipitation, with a deviation of more than 32 millimetres. Except for AFG, IND, PAK, CHN, THA, TLS, PRK, and KOR, all nations experienced annual precipitation totalling more than 1500 millimetres. Annual precipitation in IDN, MYS, PHL, PNG, SLB, and VUT was frequently greater than 2500 millimetres. The Asia-Pacific region had a significant country-wise variation in yearly precipitation ranging from ±32.73 to 280.95 mm.

5.3.2 Modelling the relationships between malaria prevalence and climatic influences

The scatter graphs of scaled malaria incidence versus temperature and precipitation in (**Figure 5.7**) demonstrate the randomness in the dataset. However, it was discovered that by classifying both predictor factors by nation, this randomness can be greatly decreased. The values of intercept and coefficients of both predictor variables were obtained by fitting the non-categorized merged data into a single GLM model. In Africa, the 95% confidence intervals for the intercept and lowest temperature are 0.053 to 0.054 and -0.065 to -0.027, respectively, with a p-value of <0.001, but values for the precipitation coefficient (95% CI: -0.018, -0.012) have much lower p-values. The AIC and BIC numbers are comparatively high, 3498.3 and 3512.6, respectively (**Table 5.1**).

Table 5.1 Model fit statistics

Models	AIC		BIC		Log Likelihood		Intercept p-value		Cook's distance threshold value	
	M-CA	M-APAC	M-CA	M-APAC	M-CA	M-APAC	M-CA	M-APAC	M-CA	M-APAC
GLM	542.90	1170.7	553.22	1180.8	-268.45	-582.33	5.43×10^{-322}	1.20×10^{-162}	0.017	0.019
GLM-G	350.99	1067.5	464.59	1178.9	-143.49	-500.76	8.93×10^{-77}	3.07×10^{-61}	0.018	0.027
GLME	-2349.1	-1292.3	-2314.7	-1258.5	1184.6	656.15	3.06×10^{-259}	3.19×10^{-51}	--	--

Model fit statistics indicating that consideration of abrupt regional- and country wise variations of the climatic factors have significantly been improved the relationship with the malaria incidence

Note: M-CA: Malaria endemic part of Central Africa; M-APAC: Malaria endemic part of Asia-Pacific region.

Table 5.2 GLM estimates of association coefficients for temperature and precipitation

Region	Country name	Code	Temperature estimates	Precipitation estimates
Africa	Angola	AGO	-0.035	-0.012
	Burundi	BDI	0.012	0.013
	Benin	BEN	0.040	0.011
	Burkina Faso	BFA	0.048	0.014
	Botswana	BWA	0.097	-0.012
	Central African Republic	CAF	0.031	0.018
	Cote d'Ivoire	CIV	0.071	0.009
	Cameroon	CMR	0.072	0.009
	Democratic Republic of Congo	COD	0.025	0.020
	Congo	COG	0.058	0.014
	Comoros	COM	0.119	0.007
	Cape Verde	CPV	-0.500	-0.033
	Ethiopia	ETH	0.043	0.030
	Gabon	GAB	0.047	0.014
	Ghana	GHA	0.078	0.015
	Guinea	GIN	0.048	0.012
	Gambia, The	GMB	0.063	0.011
	Guinea-Bissau	GNB	0.065	0.015
	Equatorial Guinea	GNQ	0.064	0.012
	Kenya	KEN	0.041	0.017
	Liberia	LBR	0.039	0.011
	Madagascar	MDG	-0.016	0.015
	Mali	MLI	0.033	0.013
	Mozambique	MOZ	0.050	0.012
	Mauritania	MRT	-0.033	0.021
	Malawi	MWI	0.060	0.015
	Namibia	NAM	0.108	0.014
	Niger	NER	0.027	0.013
	Nigeria	NGA	0.052	0.011
	Rwanda	RWA	-0.034	0.014
	Sudan	SDN	0.070	0.012
	Senegal	SEN	0.030	0.018
	Sierra Leone	SLE	0.053	0.011
	Somalia	SOM	0.101	0.012
	South Sudan	SSD	0.037	0.012
	Sao Tome and Principe	STP	0.377	0.005
	Chad	TCD	0.036	0.014
	Togo	TGO	0.062	0.014
	United Republic of Tanzania	TZA	0.049	0.014
	Uganda	UGA	0.024	0.016
	South Africa	ZAF	0.088	0.005
	Zambia	ZMB	0.021	0.010
Region	**Country name**	**Code**	**Temperature estimates**	**Precipitation estimates**
Asia-Pacific	**Afghanistan**	**AFG**	**0.005**	**0.013**

	Bangladesh	**BGD**	**-0.110**	**-0.014**
	Bhutan	**BTN**	**0.007**	**-0.067**
	China	**CHN**	**0.045**	**0.457**
	Indonesia	**IDN**	**0.076**	**-0.020**
	India	**IND**	**0.000**	**0.001**
	South Korea	**KOR**	**0.025**	**-0.011**
	Sri Lanka	**LKA**	**1.045**	**0.039**
	Myanmar	**MMR**	**0.020**	**-0.018**
	Malaysia	**MYS**	**0.674**	**0.008**
	Nepal	**NPL**	**-0.030**	**-0.025**
	Pakistan	**PAK**	**0.002**	**-0.011**
	Philippines	**PHL**	**0.250**	**-0.005**
	Papua New Guinea	**PNG**	**0.015**	**-0.007**
	North Korea	**PRK**	**0.017**	**-0.002**
	Solomon Islands	**SLB**	**-0.578**	**0.000**
	Thailand	**THA**	**0.209**	**-0.015**
	Timor	**TLS**	**0.428**	**-0.015**
	Vietnam	**VNM**	**0.316**	**-0.009**
	Vanuatu	**VUT**	**-0.113**	**0.002**

Cook's distances were also calculated to illustrate how each measurement affected the fitted response values. The expected threshold value of the cook distance is 0.005, and numerous data were discovered at anomalies beyond the threshold limit (**Figure 5.3 A**). In contrast, the Asia-Pacific region exhibits statistically significant temperature impacts ranging from -0.063 to -0.030, with a significant estimate of the intercept (0.066, 0.069) ($p < 0.001$). The estimate of the precipitation coefficient stays significant, with comparatively larger p-values. Model estimates are noticeably bad, as evidenced by extremely high AIC and BIC values (>2100). With several anomalies, Cook's threshold stays nearly the same (0.012). The country-based categorization of the dataset enhanced GLM-G fitting, as evidenced by reduced AIC and BIC values. Notably, model sizes are much better in Africa than in Asia-Pacific, where there was only a small increase (**Table 5.1**). After categorising the data, the

Cook's threshold estimates (Africa: 0.007; Asia-Pacific: 0.480) revealed a lower number of anomalies (**Figure 5.3**, **Table 5.1**).

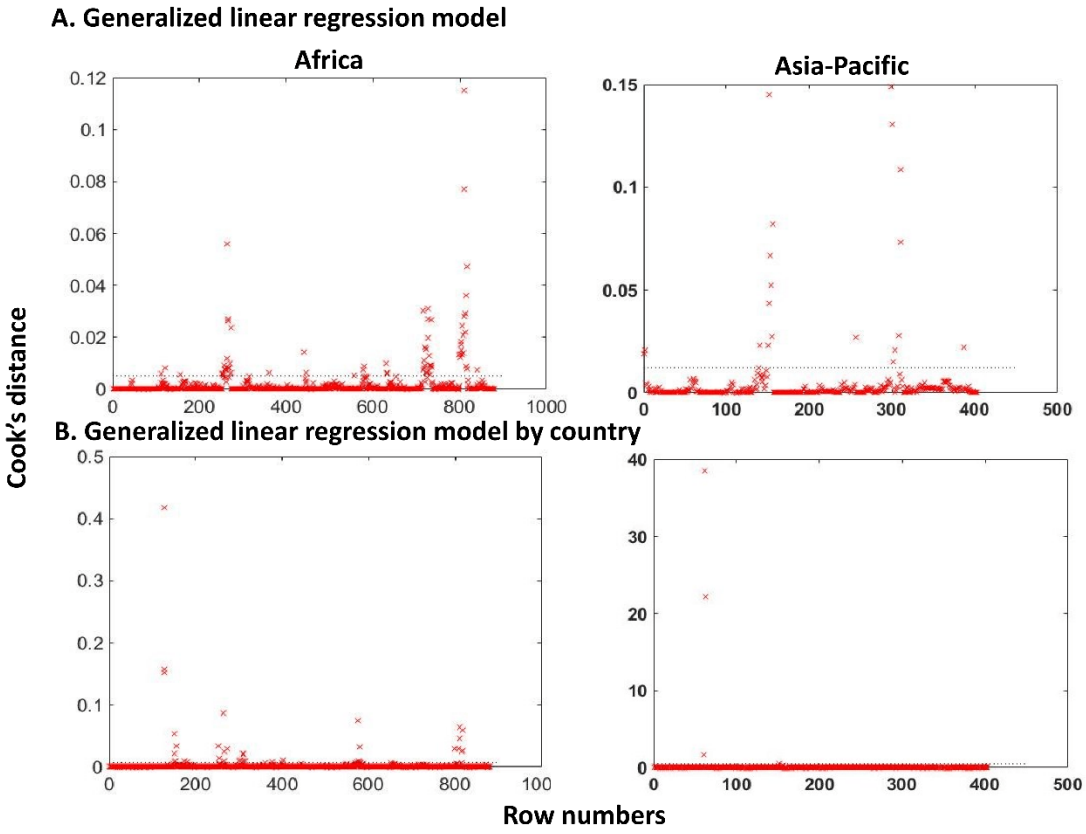


Figure 5.3 Cook's distance

To show how each measurement, shown by the red cross, affected the projected response values, Cook's distance was determined. The predicted cutoff value (dotted line) of the cook distance for Africa is 0.005 and for Asia-Pacific is 0.012. (A) GLM findings without nation classification. Cook's cutoff estimates (Africa: 0.007; Asia-Pacific: 0.480) showed a decreased number of outliers following data categorization in the GLM-G findings (B).

Fitting GLME to a country-by-country classified dataset improved the findings for both areas, as evidenced by AIC and BIC values that are more than 100 times lower than the other two models (**Table 5.1**). However, there are significant differences in intercepts and slopes for both lowest temperature and precipitation across nations.

Estimation of random effect coefficients at 95% confidence intervals reveals a significant variation in SEs across nations and between two areas. In contrast, Africa had a projected SE that was at least ten times lower than Asia-Pacific. Estimated fixed effect estimates at 95% confidence intervals revealed SEs of <0.08 in Africa and <0.28 in Asia-Pacific (**Table 5.2**, **Table 5.3**). The normal probability plot of residuals with mixed effects revealed significant variations in the impacts of two areas. It revealed that error terms are not ordinarily distributed in both areas, with Asia-Pacific errors being more skewed than Africa (**Figure 5.4**), and both regions exhibiting broad dispersions. (**Figure 5.5**) and (**Figure 5.6**) show the significant variations in the fixed and random effects of both predictor variables across nations and between two areas. The 95% CI of temperature fixed effects revealed much greater variance in Asia-Pacific, where LKA, MYS, IDN, PNG, and SLB have CI lengths greater than 1.0.

Table 5.3 GLME estimates of random effect coefficients (Alpha 0.05)

Region	Country name	Code	Temperature estimates	Precipitation estimates
Asia-Pacific	Afghanistan	AFG	-0.008	0.002
	Bangladesh	BGD	-0.003	-0.009
	Bhutan	BTN	0.007	-0.022
	China	CHN	0.031	0.199
	Indonesia	IDN	-0.004	-0.015
	India	IND	-0.007	-0.006
	South Korea	KOR	0.009	-0.001
	Sri Lanka	LKA	0.007	0.019
	Myanmar	MMR	-0.008	-0.020
	Malaysia	MYS	-0.001	-0.015
	Nepal	NPL	0.000	-0.013
	Pakistan	PAK	-0.004	-0.005
	Philippines	PHL	0.003	0.002
	Papua New Guinea	PNG	-0.009	-0.044
	North Korea	PRK	0.006	0.008
	Solomon Islands	SLB	-0.008	-0.045
	Thailand	THA	0.000	-0.013
	Timor	TLS	-0.006	-0.028
	Vietnam	VNM	0.005	0.009
	Vanuatu	VUT	-0.010	-0.004

Region	Country name	Code	Temperature estimates	Precipitation estimates
Africa	Angola	AGO	-0.006	0.002
	Burundi	BDI	-0.006	0.001
	Benin	BEN	-0.008	0.003
	Burkina Faso	BFA	-0.005	0.004
	Botswana	BWA	0.016	-0.021
	Central African Republic	CAF	-0.004	0.004
	Cote d'Ivoire	CIV	-0.005	0.003
	Cameroon	CMR	-0.002	0.004
	Democratic Republic of Congo	COD	-0.003	0.004
	Congo	COG	-0.001	0.003
	Comoros	COM	0.007	-0.005
	Cape Verde	CPV	-0.004	-0.029
	Ethiopia	ETH	0.008	0.002
	Gabon	GAB	-0.001	0.002
	Ghana	GHA	0.001	0.004
	Guinea	GIN	-0.004	0.004
	Gambia	GMB	-0.003	0.002
	Guinea-Bissau	GNB	0.005	0.001
	Equatorial Guinea	GNQ	-0.005	0.002
	Kenya	KEN	0.010	0.001
	Liberia	LBR	-0.005	0.004
	Madagascar	MDG	0.002	-0.003
	Mali	MLI	-0.006	0.003
	Mozambique	MOZ	-0.005	0.003
	Mauritania	MRT	0.008	-0.001
	Malawi	MWI	0.000	0.003
	Namibia	NAM	0.015	-0.005
	Niger	NER	-0.007	0.002
	Nigeria	NGA	-0.004	0.004
	Rwanda	RWA	-0.009	0.000
	Sudan	SDN	0.008	-0.002
	Senegal	SEN	0.009	0.001
	Sierra Leone	SLE	-0.004	0.004
	Somalia	SOM	0.005	-0.001
	South Sudan	SSD	-0.004	0.003
	Sao Tome and Principe	STP	0.000	-0.004
	Chad	TCD	-0.002	0.002
	Togo	TGO	-0.002	0.004
	United Republic of Tanzania	TZA	0.001	0.002
	Uganda	UGA	-0.005	0.003
	South Africa	ZAF	0.017	-0.016
	Zambia	ZMB	-0.005	0.001

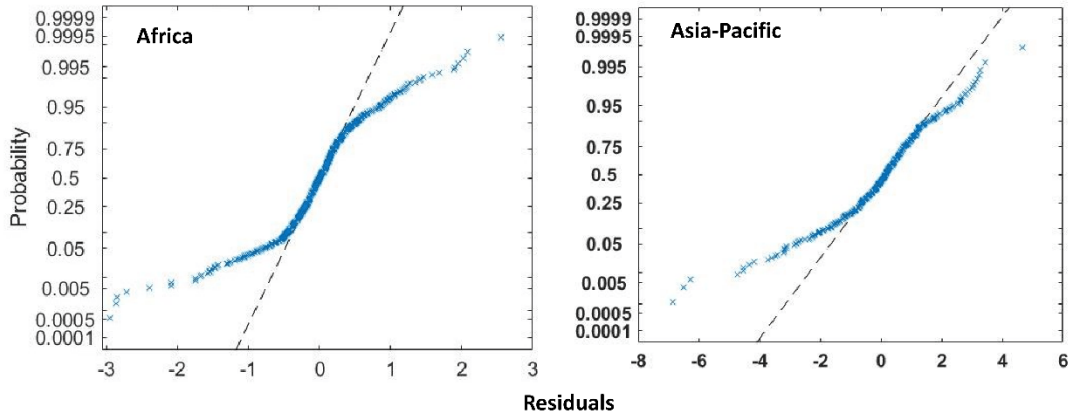


Figure 5.4 The normal probability plot of residual results

The residual findings from generalised mixed effect models were shown using a normal probability distribution, which indicated notable differences in the affects between Africa and the Asia-Pacific region. It showed that errors are not typically distributed in both regions, with errors being more skewed in the Asia-Pacific region compared to Africa.

In Africa, all nations had nearly identical fixed temperature impacts, with STP (0.38) having a slightly higher number (**Figure 5.5 A**). Precipitation fixed effects have maintained a similar pattern for Africa, with nearly the same degree of variance across nations as demonstrated by the 95% CI ranges (**Figure 5.5 B**). However, Asia-Pacific nations revealed different fixed effects of precipitation, with CHN having at least twice the value of CI length and PAK and AFG having ten times the value. Surprisingly, the random impacts of temperature and precipitation variance revealed opposing trends. While random temperature impacts in Africa vary greatly across nations, the Asia-Pacific region stays very similar (**Figure 5.6 A**).

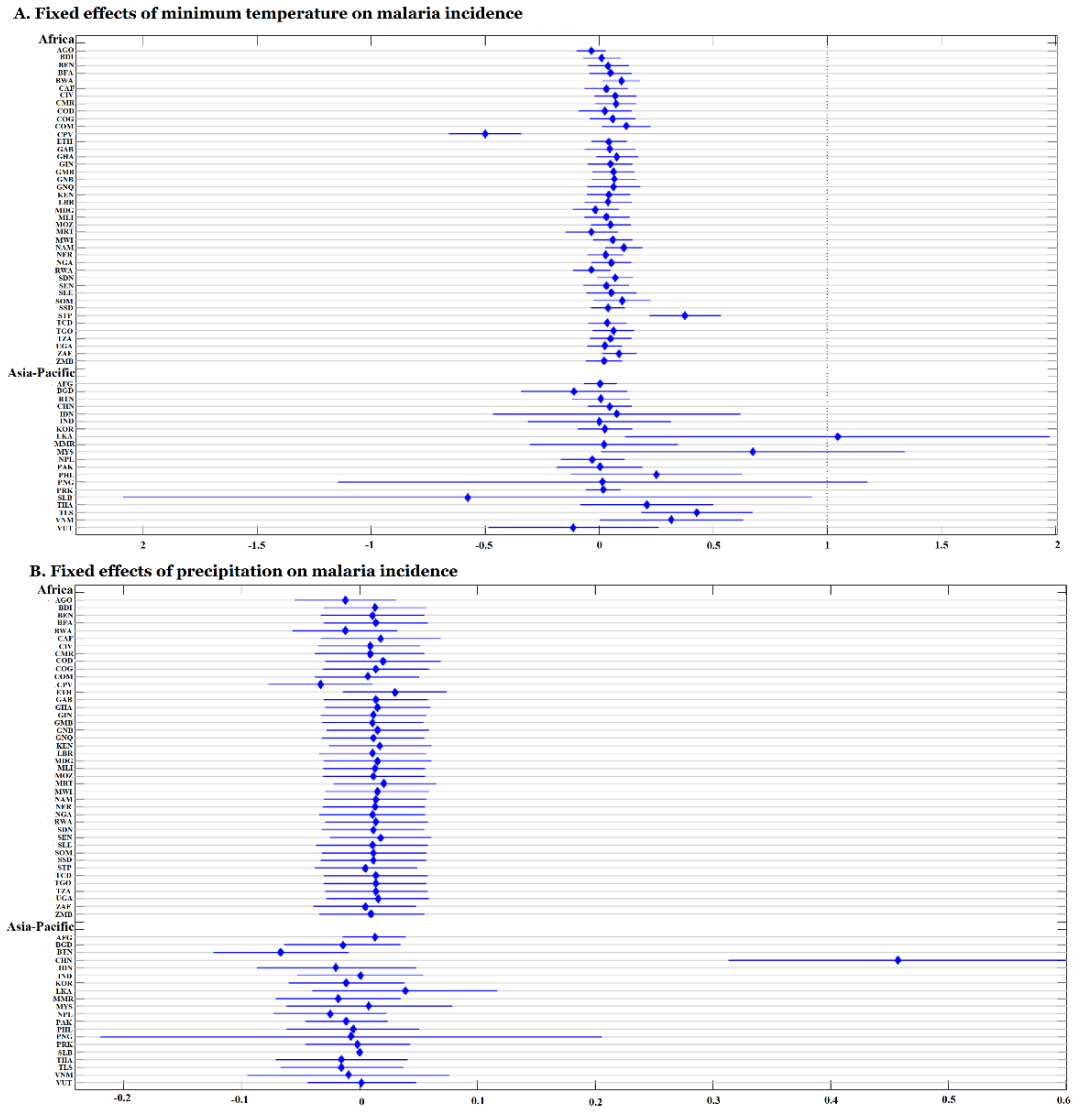


Figure 5.5 The forest Plot showing fixed effects.

The forest plot shows the substantial variations in both predictor variables' fixed effects across the countries and regions of Asia-Pacific and Africa. (A) The Asia-Pacific region showed substantially larger heterogeneity in the 95% CI of temperature fixed effects. A slightly larger value was reported in STP (0.377), COM (0.119), NAM (0.108), and SOM (0.101). (B) Fixed impacts of precipitation have maintained the similar pattern for Africa, with virtually the same degree of variance among nations as seen in the 95%CI ranges. However, Asia-Pacific countries showed significant fixed impacts of precipitation, with PAK and AFG showing at least five times lower CI values while CHN and PNG showing at least two times higher CI values.

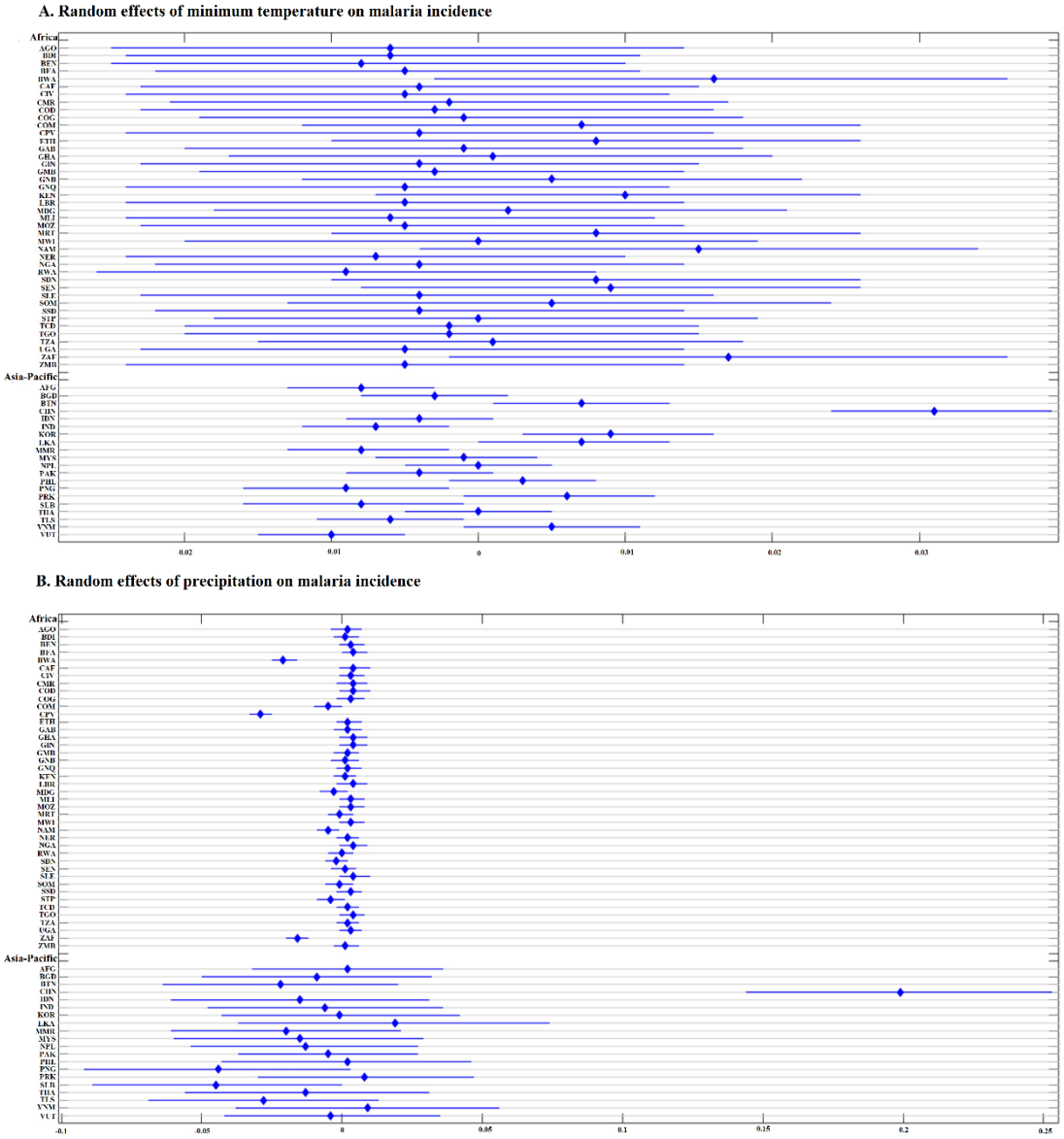


Figure 5.6 The forest plot showing random effects

A forest plot illustrating the unpredictable effects of changing precipitation and weather. (A) The Asia-Pacific region is mostly unaffected by random temperature changes, but the effects in Africa vary substantially between nations. (B) In contrast, the impacts of random precipitation in Africa were quite similar, but the Asia-Pacific region showed notable regional variations. The predicted range length of random temperature impacts (95% CI) in Africa is 0.03-0.04, whereas in Asia-Pacific it is 0.01-0.015. The estimated range length for random precipitation impacts, however, is 0.063-0.111 in Asia-Pacific and 0.007-0.012 in Africa.

Random precipitation impacts in Africa, on the other hand, have stayed very similar, whereas Asia-Pacific has shown substantial variations across nations (**Figure 5.6 B**). CHN and SLB stood out as notable exceptions to this trend. The calculated length of the range of random temperature impacts (95% CI) in Africa is 0.033-0.040, while it is 0.010-0.015 in Asia-Pacific. However, projections of the CI length for the random precipitation impact in Africa range from 0.007 to 0.012, while in Asia-Pacific it ranges from 0.064-0.111 (**Table 5.3**).

Spatial patterns of malaria correlations with annual lowest temperature and precipitation are created across all 62 countries by grouping or clustering countries with comparable associations (**Figure 5.7**).

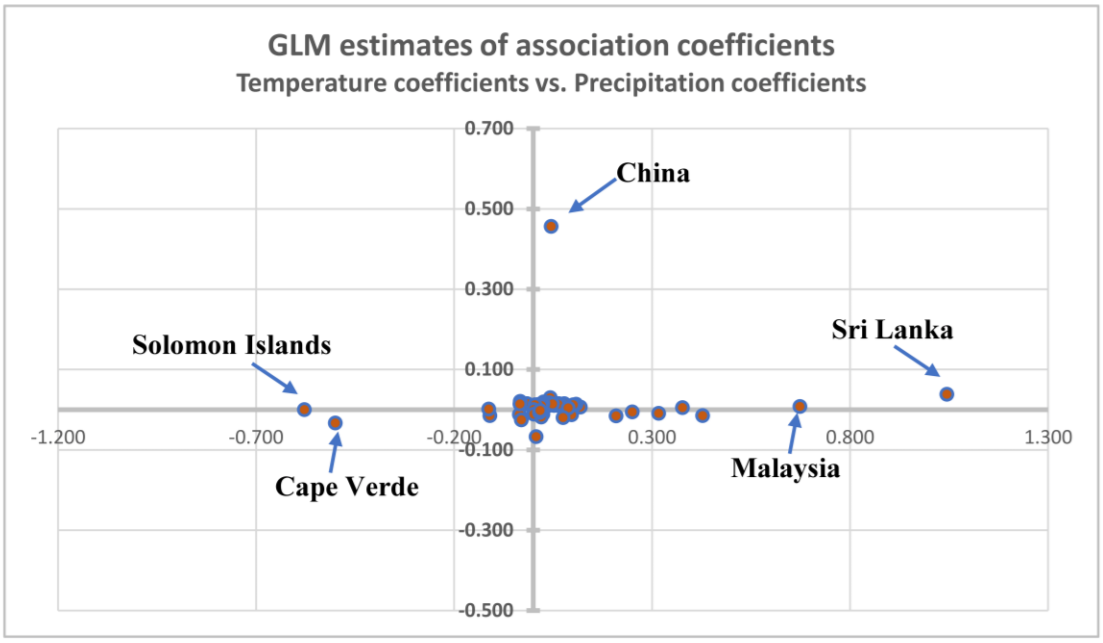


Figure 5.7 Scatter graphs of GLM-G estimates of temperature and precipitation factors

GLM-G estimations of the effects of temperature and precipitation on the frequency of malaria are plotted in scatter graphs. The arrow points to the regions that have distinct connection associations.

For the country-wise GLM-G fixed values of temperature and precipitation factors, we used an agglomerative clustering algorithm with a “Euclidian” distance measure. This approach iteratively combines clusters along the bottom-up structure (**Figure 5.8**) using a linking function “ward” that reduces cluster variance. Outlier nations are plainly visible in this spatial pattern **(Figure 5.9**). MYS and LKA are in cluster “0”, CPV and SLB are in cluster “2”, STP, PHL, THA, TLS, and VNM are in cluster “3”, and CHN is the only one in cluster “4”. Cluster “1” includes all other nations. It also shows that, with very few instances, African nations have a very similar malaria association.

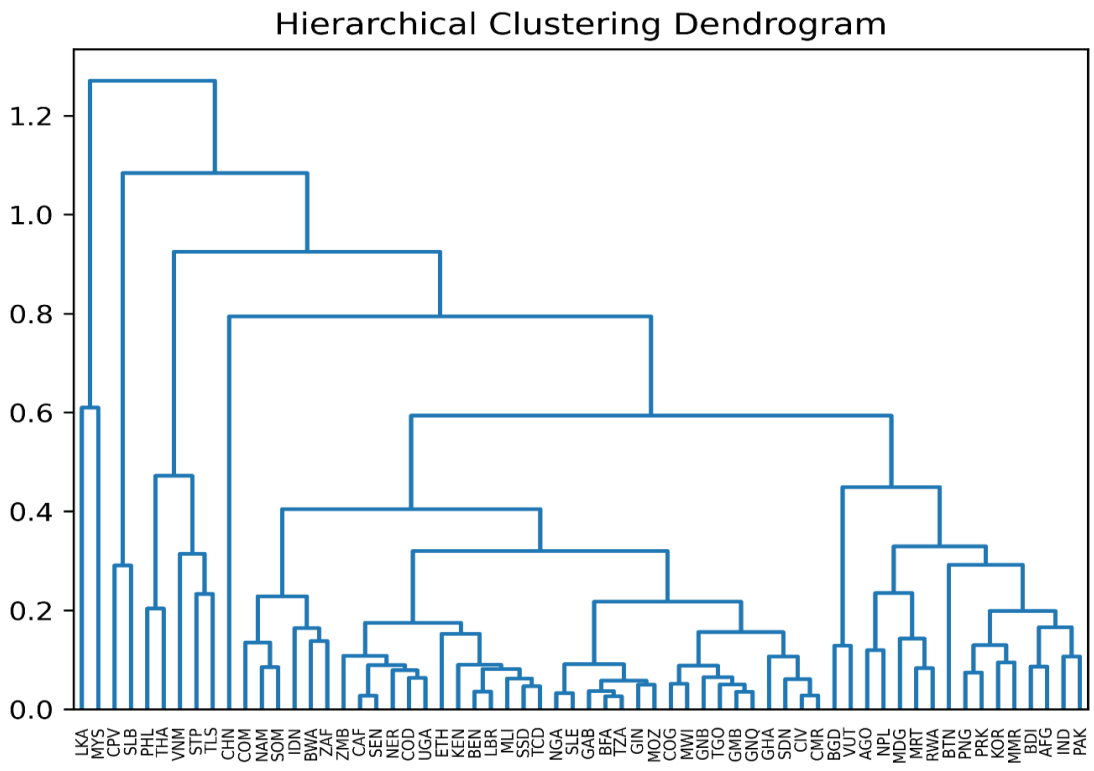


Figure 5.8 Agglomerative clustering

The “Euclidian” distance metric was utilised in conjunction with the agglomerative clustering approach to determine the country-level GLM-G fixed values of the temperature and precipitation coefficients. This method use an iterative linking function called “ward” to unite the clusters along the bottom-up structure (i.e., dendrogram) while lowering the variance of the individual clusters.

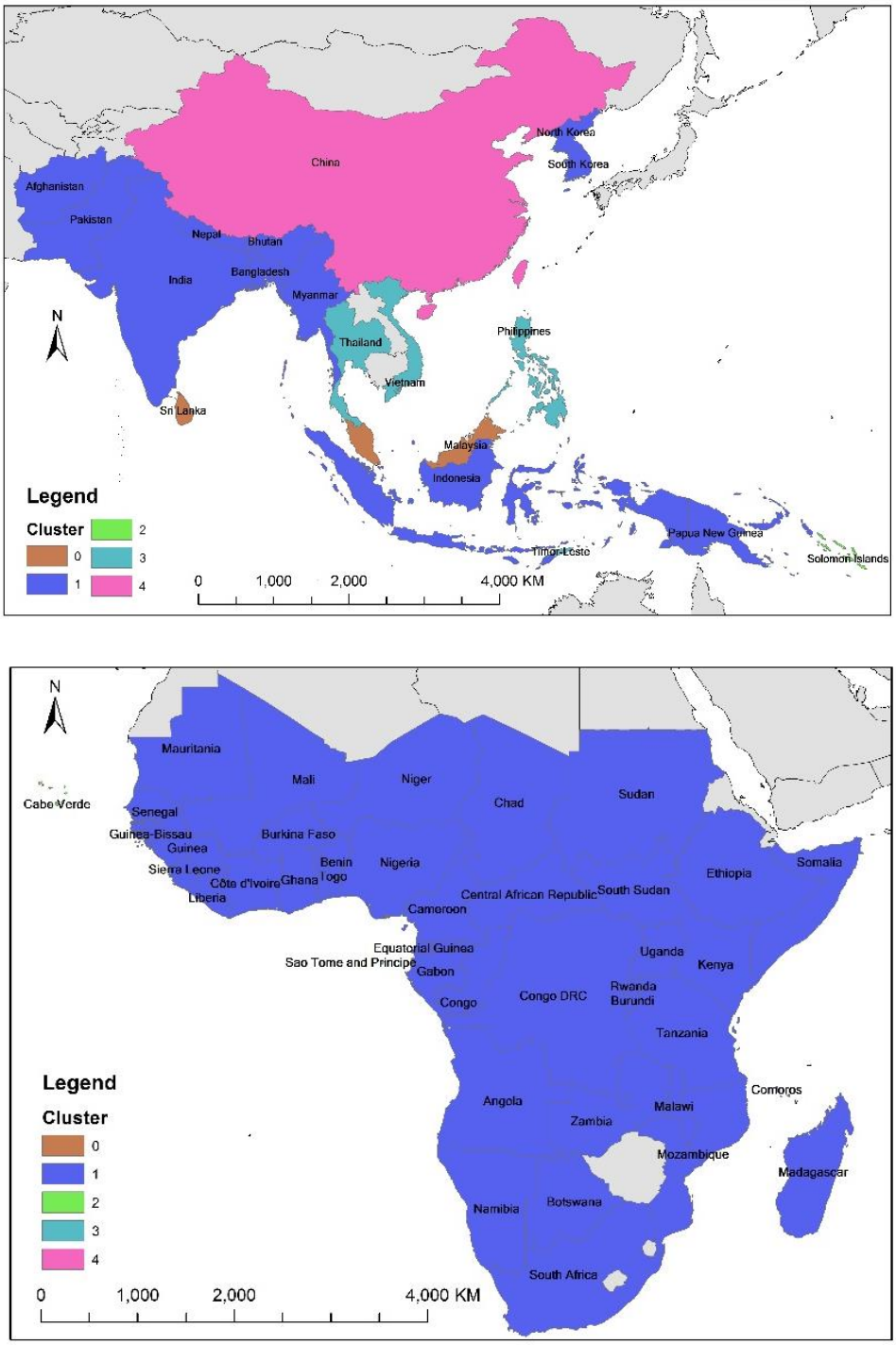


Figure 5.9 Agglomerative clustering

All 62 countries (42 in Africa and 20 in Asia-Pacific) are grouped together into five geographical zones (i.e., 0, 1, 2, 3, and 4) that have malarial association correlations with the yearly lowest temperatures and precipitation that are quite similar to one another. It unmistakably shows how similar African nations are to those in the Asia-Pacific. STP and CPV have distinctive linkages in Africa, whereas CHN, MYS, and LKA have unique associations in the Asia-Pacific area.

5.4 Discussion

Malaria has had a significant impact on global health by raising the huge health burden, and has been found to be substantially higher in tropical and temperate regions, but very few attempts have been made to measure and describe the region-specific differential patterns in malaria burden. The total incidence and health burden of malaria have decreased considerably in malaria-endemic areas as a result of large-scale adoption and successful application of disease-related intervention initiatives. The current study revealed intra- and inter-regional variations in malaria incidences in Africa and Asia-Pacific from 2001 to 2020. (**Figure 5.1**). The variance was noted at both the regional and national levels, and the findings revealed a remarkable decrease in malaria prevalence in Asia-Pacific, but the amount was found to be overlapping and considerably greater in Africa. Several studies have found that favourable climate conditions, landscape utilisation, drug resistance, and deliberate malaria intervention strategies (e.g., interruption of residual insecticides application) are the primary factors influencing malaria spread and transmission, posing serious challenges to effective control and elimination in endemic[161-164]. According to the WHO World Malaria Report 2021, malaria incidence per 1000 people at risk decreased from 81 in 2000 to 59 in 2015 and 56 in 2019, but rose to 59 in 2020. This rise in 2020 has been linked to the interruption of critical malaria services during the COVID-19 pandemic. During the COVID-19 period, malaria services in high-burden to high-impact countries (HBHI) experienced a slew of disruptions, including a shattered supply chain for health commodities and increased costs for purchasing, shipping, and distribution, which fluctuated over time as COVID-19 regulations changed. Because of the high fluctuating effect, quantifying the degrees of disruption to malaria

networks has proven challenging. To deal with such circumstances and uncertainty about the number of cases, WHO developed three distinct techniques for estimating malaria cases based on area and degree of malaria transmission. Based on this high-quality dataset, the current study aimed to assess the relationships between annual temperature and precipitation variability and the incidence of malaria in two regions (**Figure 5.3**, **Figure 5.4**, **Figure 5.5**). Such an assessment will undoubtedly aid in determining intra- and inter-regional differences and disease risk stratification, and it is critical to identify the magnitude that provides opportunities to control the responsible factors, particularly in malaria-endemic areas.

According to recent research, the total threat of malaria in afflicted areas is largely limited due to favourable climatic variables, vector and parasite behaviour, and transmission levels [165, 166]. Temperature and rainfall variability are major determinants of malaria spread and transmission, significantly increasing overall malaria incidence and disease burden risks in affected regions due to the linkage with key climate change factors[167-172]. Thus, the current research found that temperature changes are the primary causes of such overlapping malaria prevalence, favouring high local transmissions in the impacted areas. According to the calculated association coefficients, both yearly lowest temperature and precipitation variations can have a beneficial impact on malaria prevalence in Africa (**Figure 5.6**, **Figure 5.9**). The GLME-estimated random effect factors represent the intrinsic randomness in temperature and precipitation variance. These results, however, may be attributed to the beneficial impact of good climate conditions, increases in temperature and rainfall, and physiological suitability in the spatiotemporal patterns in malaria-affected areas. According to [173], the trend of malaria has increased with rising

temperature, and transmission declines with precipitation and rises during the arid seasons in China. Similarly, several Asia-Pacific nations, including BGD, IND, IDN, MMR, MYS, PAK, PNG, SLB, THA, TLS, and VUT, demonstrated comparatively significant unfavourable random impacts of temperature and precipitation changes. However, in general, African nations exhibited a mixed form of random effect; while annual temperature variation primarily caused negative random effects, annual precipitation variation primarily caused positive random effects. It suggests that temperature has a significant impact on disease spread in both areas. Similar to this finding, newer studies have found a positive correlation between malaria prevalence and temperature and precipitation, with the exception of a few years in which a negative association with rainfall was noted in African areas [174] Agglomerative clustering of all 62 nations results in five spatial zones, each with a very comparable malarial association relationship with yearly temperature and precipitation. It definitely demonstrates that African nations are extremely similar to Asia-Pacific countries. While CHN and SLB have unique associations in Asia-Pacific, STP and CPV have distinct associations in Africa **(Figure 5.9**). Such unique correlations in outlier countries can be explained by yearly malaria rates that have stayed relatively low or high over the last two decades or have shifted dramatically in consecutive years. For example, except in 2000 (1.293) and 2017 (3.027), CPV in Africa regularly demonstrated comparatively low malaria incidence (<1). In parallel, SLB in the Asia-Pacific showed comparatively high malaria prevalence (>100) until 2011, with the maximum level of 681.26 in 2001, and then it suddenly decreased to 52.39 and 66.67 in 2014 and 2015, respectively, before increasing in following years to 167.67 in 2020. In comparison to Africa, Asia-Pacific nations showed a wide range of correlations between malaria incidence, yearly minimum temperature, and

precipitation. These variations could be related to interannual variations in precipitation caused by shifts in the Asian monsoon circulation. The Asia-Pacific region has a normal monsoon environment, with heavy rain in the summer and little rain in the winter. In most areas of the region, the summer monsoon accounts for nearly 75% of total yearly rainfall. However, significant differences in the start and length of the summer monsoon across nations are obvious and important for disease dynamics and epidemiology. Such variations are mirrored in our findings, with temperature and precipitation having varying impacts on malaria prevalence across Asia-Pacific countries (**Figure 5.5)**. The determination of random factors is also influenced by the precipitation's inherent randomness **(Figure 5.6**). This thought has increased the importance of the links between malaria incidence and climatic variables. According to other studies, climate change with rising temperature can raise the possibility for a malaria outbreak in both areas of nations that are extremely prone to the illness[175-177]. In nations where malaria has been eliminated or managed, rising temperatures can impact the reintroduction or raise the incidence of the disease. As a result, it is clear that monitoring and preparation in those emerging nations must be prioritised, as they must balance several conflicting interests for limited resources, many of which are related to healthcare services.

Despite the fact that numerous studies have been conducted to investigate the relationship between malaria incidence, climatic factors, and potential climate change scenarios[178-181], it is noted to be highly complex due to its interdependence on spatiotemporal scales, socioeconomic factors, access to health services, variable measures of interventions, and mosquito biology. Disentangling the relationships between malaria incidence and climatic variables by excluding the impacts of all

possible covariates is extremely challenging in terms of developing a suitable approach, model, and scale, as well as interpreting the findings. The sensitivity of malaria to climate, on the other hand, continues to generate significant worry about the consequences of climate change on future disease dynamics. The issue of malaria vectors migrating from their native habitats to infiltrate new zones is of special worry. As a result, model-based studies incorporating rainfall, temperature, and other climatic variables remain important for methodically addressing these issues and providing critical inputs for developing and executing successful intervention strategies.

5.5 Summary

1. In this chapter Time Series data analysis of Malaria Incidence in Asia-Pacific region and Africa was conducted.

2. The present study showed high intra-and inter regional differences in malaria incidence, with almost strictly decadal decreasing trends in the Asia-Pacific region and a kind of mixed trends in Africa.

3. Furthermore, it showed that malaria incidences are significantly associated with the annual minimum temperature and precipitation, although there are high variations across the countries in Africa and Asia-Pacific region.

4. In contrast, most Asia-Pacific countries showed negative precipitation effects.

5. It found that most Asia-Pacific countries hold negative random effects caused by the both temperature and precipitation variations. In Africa, while

temperature variation frequently caused negative random effects, precipitation variation caused positive effects.

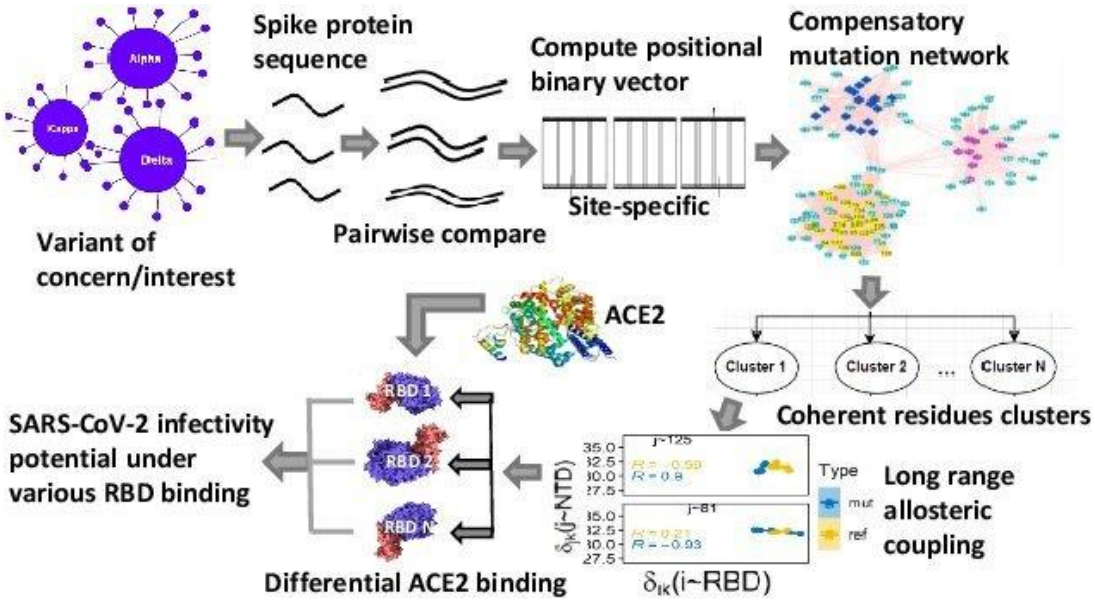
6. This study signifies the association between malaria incidence and climatic factors and its intra-and inter regional differences at large spatial scale.

"Data analysis is the bottleneck for making progress in proteomics" … Ruedi Aebersold

Chapter 6

6. Protein data analysis and applications

Graphical Abstract



Das JK, **Thakuri B**, MohanKumar K, Roy S, Sljoka A, Sun GQ, Chakraborty A. Mutation-Induced Long-Range Allosteric Interactions in the Spike Protein Determine the Infectivity of SARS-CoV-2 Emerging Variants. *ACS Omega. 2021 Nov 10;6(46):31312-31327. doi: 10.1021/acsomega.1c05155. PMID: 34805715; PMCID: PMC8592041*.

6.1 Introduction

Global attempts are being made to create vaccines and antiviral medications that are more effective in order to combat the continuing COVID-19 pandemic, which has killed 4,170,155 people as of July 2021. The fast rise of numerous SARS-CoV-2 variants, the cause of COVID-19, is severely impeding these efforts[182-184]. There is mounting evidence that SARS-CoV-2 variants that altered the antigenic profile can subvert immune responses and lessen the antigen-neutralizing effects of antibodies[185-187]. Recent research also suggests that convalescent plasma and mAb therapies may be exerting a selective impact on the evolution of novel variants[188-190]. Long-term viral shedders may aid in the random appearance of more severely mutated variants under such selective pressure. The WHO technical advisory group has identified a number of variations of concerns (VOC) or variations of interests (VOI) that are in circulation worldwide (**Table 6.1**). These variations call for immediate consideration in order to better define control measures. These variants are known to cause substantial community transmission or numerous COVID-19 clusters in various nations, with rising relative frequency and rising case numbers over time, with the mentioned suite of mutations. Here, we have focused on the important Spike RBD mutations E484K, K417N, L452Q, L452R, N501Y, and T478K that are frequently found in the VOI/VOC forms (**Table 6.1**). These RBD mutations are known to distinguish and define a number of new variants. Changes in several viral traits, such as transmissibility, illness severity, immune escape, and diagnostic or therapeutic escape, help identify VOC/VOIs with these mutations[184, 185, 191, 192]. One such widely used VOC that first emerged in India in late 2020 is the delta variant, which has a very high transmissibility. Lineage B.1.617.2 is the name given

to this variant, which is distinguished by a number of spike changes. Of these, the spike protein's T478K, K417N changes set it apart from other VOIs in a distinctive way. Although new studies have documented comparative evaluations of variant characteristics, it is still unknown how such mutations impact virus characteristics with apparently unrelated sporadic mutations. By attaching its prefusion-form spike protein (S) to the human angiotensin converting enzyme-2 (ACE2), which is highly expressed on the surface of lungs, heart, kidney, and intestine cells, coronaviruses (CoVs) can identify and penetrate into human host cells. Most vaccine research and therapeutic efforts centre on the S-protein to prevent this crucial entry mechanism and following infection[193]. Each protomer of the trimerized S-protein consists of two subunits: the amino (N)-terminal S1 and the carboxyl (C)-terminal S2[194]. Cellular protease furin cleaves the S1/S2 junctions at Arg685-Ser686 during proteolytic processing and subsequent membrane union. The prefusion interactions between the component S1 and the human target protein ACE2 are its primary function[195]. It is made up of the receptor-binding domain (RBD), two highly conserved segments (SD1 and SD2), and an N-terminal domain (NTD). A variety of RBD conformational configurations have been shown in various investigations, varying between the RBD-up position that is favourable for ACE2 binding and the RBD-down position that is comparatively refractory to receptor binding[196-198]. The majority of RBD and NTD-based VOC/VOI variants that distinguish one from the others have an impact on interactions with ACE2[185]. For instance, the L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, H655Y, and T1027I amino-acid changes, which are shown to decrease antibody neutralisation, are characteristic of the P.1 lineage, which was first discovered in Brazil. The impact of the NTD mutations on the associations with the ACE2 that attaches to the RBD is unclear, though. Here, we investigate whether

any grouped or partial NTD forms interact allosterically with the six RBD mutational sites identified as VOC/VOI[199].

In addition to structural modifications, kinetics are also affected by effector dependence, which can cause long-range allosteric disturbances to spread. Experiments using statistical thermodynamics demonstrate that ligand binding can produce free energies through beneficial interactions. Potential changes in macromolecular thermal variations brought on by ligand binding encompass a variety of dynamic interactions, from randomly occurring local anharmonic movements of molecular regions to highly correlated, low-frequency normal mode vibrations. The entropy effect is mainly responsible for this type of dynamic allostery. [200-202]. When long-range allosteric disturbances spread, structural alterations are not required because of the effector dependent modulations of structural dynamics. Comparative studies of the apo- and effector-bound states help identify the endpoints of this long-range allostery. The distal dynamic region of the protein structure is connected by a network of AA residues that carry out such allosteric propagations of disruptions[203]. Understanding and forecasting the various effects of the VOIs depend on the discovery of this AA residue network, which supports mutation-induced dynamic allosteric modulation. In this study, we investigated a variety of VOC/VOI mutation networks differentially connecting NTD and RBD of the SARS-CoV-2 spike protein using computationally predicted chemical shifts data of ^{1}H and ^{15}N, calculated based on SHIFTX2[204] that combines ensemble machine learning methods with sequence alignment-based methods. We have developed an integrated approach using sequence, structure, and chemical shift data to deduce such long-range allosteric connections modulated by SARS-CoV-2 VOC/VOI RBD mutations, in

addition to the well-studied chemical shift covariance analysis (CHESCA)[205]. Cytoscape plugins[206] were used to derive highly linked sub-graphs of susceptible mutation sites discovered from AA sequences of Spike variants using the Molecular Complex Detection (MCODE) technique[207]. To determine the impacts of the identified RBD mutations, the chemical shift projection analysis (CHESPA)[208] was applied to spike chemical shift data from both mutant and non-mutated samples. Then, using the Protein Binding Energy Prediction (PRODIGY)[209] and PDBePISA webservers[210], the binding affinity for RBD and ACE2, the dissociation constant, the H-bond, and the salt-bridges were calculated. It demonstrates that highly linked mutation sites at NTD are divided into distinct groups using various combinations of secondary structural elements, and that this creates a potent allosteric connection with the mutational site at RBD. The delta variant with the RBD mutation K417N in particular exhibits a powerful long-range modulated allostery, leading to increased interactions with ACE2. These findings offer crucial information for the anticipated creation of allosteric modulators that prevent interactions with ACE2 and thereby prevent viral entry.

Table 6.1 SARS-CoV-2 variant of concern (VOC) and variant of interest (VOI)

	WHO level	Pango linage	spike protein substitutions	Name (next strain	First Detected	Remarks
A) Variant of Concern						
	α	B.1.1.7	Spike: 69del, 70del, 144del, (E484K[a]), (S494P[a]), N501Y, A570D, D614G, P681H, T716I, S982A, D1118H (K1191Na)	20I/501Y.V1	United Kingdom, Sep 2020	increase in transmissibility or detrimental change in COVID-19 epidemiology; OR increase in virulence or change in clinical disease presentation; OR decrease in effectiveness of public health and social measures or available diagnostics, vaccines, therapeutics
	ß	B.1.351	Spike: D80A, D215G, 241del, 242del, 243del, E417K, E484K, N501Y, D614G, A701V	20H/501.V2	South Africa, May 2020	
	Delta[b]	B.1.617.2	Spike: T19R, (G142D), 156del, 157del, R158G, (K417N[a]) L452R, T478K, D614G, P681R, D950N	20A/S:478K	India, Oct 2020	
	γ	P.1	Spike: L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I	20J/501Y.V3	Japan/Brazil, Nov 2020	
(B) Variant of Interest						
	Iota	B.1.526	L5F, (D80G[a]), T95I, (Y144-[a]), (F157S[a]), D253G, (L452R[a]), (S477N[a]), E484K, D614G, A701V, (T859N[a]), (D950H[a]), (Q957R[a])	20C/S:484K	USA, Nov 2020	SARS-CoV-2 with genetic changes that are predicted or known to affect virus characteristics such as transmissibility, disease severity, immune escape, diagnostic or therapeutic escape; AND identified to cause significant community transmission or multiple COVID-19 clusters, in multiple countries with increasing relative prevalence alongside increasing numbers of cases over time, or other apparent epidemiological impacts to suggest an emerging risk to global public health
	Kappa	B.1.617.1	Spike: (T95I), G142D, E154K, **L452R**, **E484Q**, D614G, P681R, Q1071H	20A/S:154K	India, Dec 2020	
	Eta	B.1.525	A67V, 69del, 70del, 144del, **E484K**, D614G, Q677H, F888L	20A/S:484K	UK and Nigeria, Dec 2020	
	Lambda	C.37	G75V, T76I, δ246-252, L452Q, F490S, D614G, and T859N	21G	Peru, Dec 2020	

6.2 Results and Discussion

6.2.1 Network of compensatory mutations in SARS-CoV-2 spike types

In the context of a deleterious mutation, compensatory mutation improves survival; otherwise, it is neutral or detrimental. It typically occurs over gene sequences non-randomly and is more likely than predicted by chance close to the location of the real harmful mutation. We gathered SARS-CoV-2 spike (S) protein sequence variations from NCBI database in order to find a group of such compensatory mutations that are frequently added after a single mutation in biological sequences and are in charge of preserving conformational and functional stability. For the purpose of the upcoming alignment study, we have collected a total of about 91 thousand sequence samples, with each sequence variant of length 1273 being noticed in at least three distinct samples ($n \geq 3$). (Table S1) https://pubs.acs.org/doi/full/10.1021/acsomega.1c05155.

Among all the SARS-CoV-2 spike protein sequence samples, we discovered 1784 distinct sequences (N = 1784) that were grouped together. We analysed all conceivable pair-wise sequence variants in each location of the sequence as opposed to reference-based comparison. As a result, a population of binary sequence variants of size $M\binom{1784}{2}$ (=1590436 sequences) is created, with 0 denoting no substitution and 1 denoting a single AA substitution. (**Figure 6.1**)

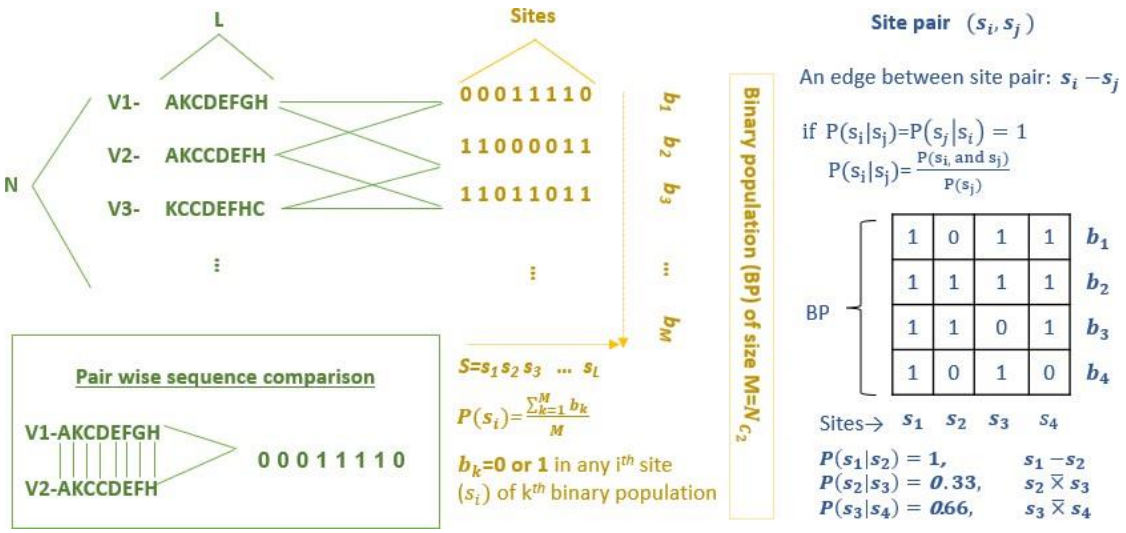


Figure 6.1 A probability scheme for identifying compensatory changes based on amino-acid (AA) sequences

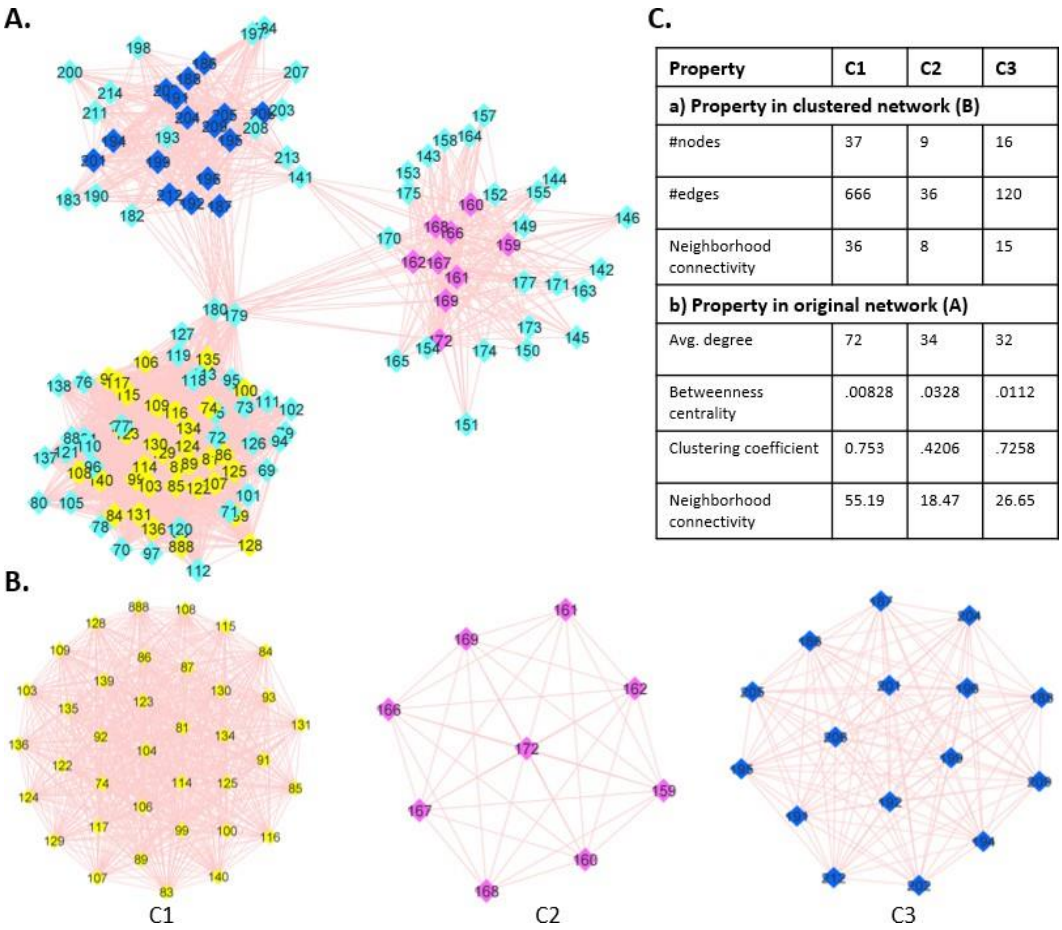
All SARS-CoV-2 variants' mixed AA unique sequences of the S protein are listed in the left panel. These sequences were then compared pairwise to produce the binary population of the sequence seen in the middle panel (N=1784, L=1273, M=1590436); "1" denotes an AA substitution and "0" denotes an absence. The right column illustrates how conditional probability was used to evaluate two mutation sites across the S-binary population (BP). If both the conditional chance P (si|sj) and P (sj|si) are 1, then the mutation site i is compensating for the site j, or vice versa.

In the spike protein, this resulted in a cumulative tally of 618 mutation sites (about 50% locations). With these residual sites found, we calculated pair-wise joint and conditional probabilities over the binary population to look at compensatory links between these locations (Methods). We chose all the couples with the greatest equal conditional probability of 1, inferring their powerful compensatory effects, in a pair-wise analysis of the mutation positions. Out of a total of $\binom{618}{2}$ (=190653) pair-wise

mutation sites, only 152 nucleotide sites and 2671 pair-wise links were found (Table S2).https://pubs.acs.org/doi/full/10.1021/acsomega.1c05155. So, it offered an undirected, unweighted network with each 152 mutation site serving as a network node and each 2671 paired link serving as an edge. By eliminating all isolated nodes, we have created a component (i.e., a maximally connected network), which ultimately refers to a compensatory mutation network of SARS-CoV-2 spike protein (**Figure 6.2** A). We have observed that the complete network is contained within the NTD (resi 13-305) of the spike protein, which is similar to the typical feature of compensatory mutations that tend to appear more frequently in specific regions of the protein. With a multimodal degree distribution, it is a fully linked mutation network with 90% of mutation sites (136 nodes out of 152 and 2660 edges out of 2671) (**Figure 6.8**).

The SARS-CoV-2 Spike compensatory mutation network (**Figure 6.2** B) exhibited aggregation of several network nodes (used MCODE[207]) around the selected multimodal degree distributions of the entire network. We have identified three of these groups, denoted by the subnetworks C1, C2, and C3, each of which has distinctive network characteristics (**Figure 6.2** C). Furthermore, the only three nodes—resi 149, 179, and 180—that link these three subnetworks to the rest of the compensating mutation network are deleted, completely severing their connections. Nearly straight spike protein basic structure sites are present in each of the clusters (C1: resi 74-140; C2: resi 159-172; C3: 186-212) (Table S2). https://pubs.acs.org/doi/full/10.1021/acsomega.1c05155. These nodes indicated beta strands in the NTD of the S protein when they were projected onto the 3D structure of the S protein (**Figure 6.3** A) (**Figure 6.3** B). A beta sheet with three strands makes up C1, a beta sheet with one strand makes up C2, and a beta sheet with two strands makes up C3.

Despite being relatively near to one another in 3D space, they spread allosteric effects to various RBD sites differently depending on the RBD mutation, which affects how well they attach to the human receptor ACE2 (described in the following sections). Considering the high degree of structural plasticity of the NTD and RBD domains (**Figure 6.3** C), there may be a large number of additional mutational combos that call for compensatory changes, are consistent with high viral fitness, and may help the immune system flee effectively. For instance, a recent investigation revealed that N439K compensated for the RBM mutation K417V, which would otherwise reduce receptor binding affinity, and that several mAbs were more susceptible to these mutations when combined than when they were present separately.



C.

Property	C1	C2	C3
a) Property in clustered network (B)			
#nodes	37	9	16
#edges	666	36	120
Neighborhood connectivity	36	8	15
b) Property in original network (A)			
Avg. degree	72	34	32
Betweenness centrality	.00828	.0328	.0112
Clustering coefficient	0.753	.4206	.7258
Neighborhood connectivity	55.19	18.47	26.65

Figure 6.2 The "AA compensatory mutation network (CMN)" for the SARS-CoV-2 spike protein

Colored nodes indicate the tightly linked mutation sites, and edges signify compensatory connections. (A) Visual depiction of the CMN-a connected, undirected network fully present in the N-terminal domain of the S1-unit of a spike protomer, which is divided into three parts visually and connected by nodes 141, 179, and 180; (B) Three network clusters: C1 (yellow), C2 (pink), and C3 (blue); produced by Cytoscape's molecular complex discovery (MCODE) method; (C) a list of various quantitative traits that set the groups C1, C2, and C3 apart from one another.

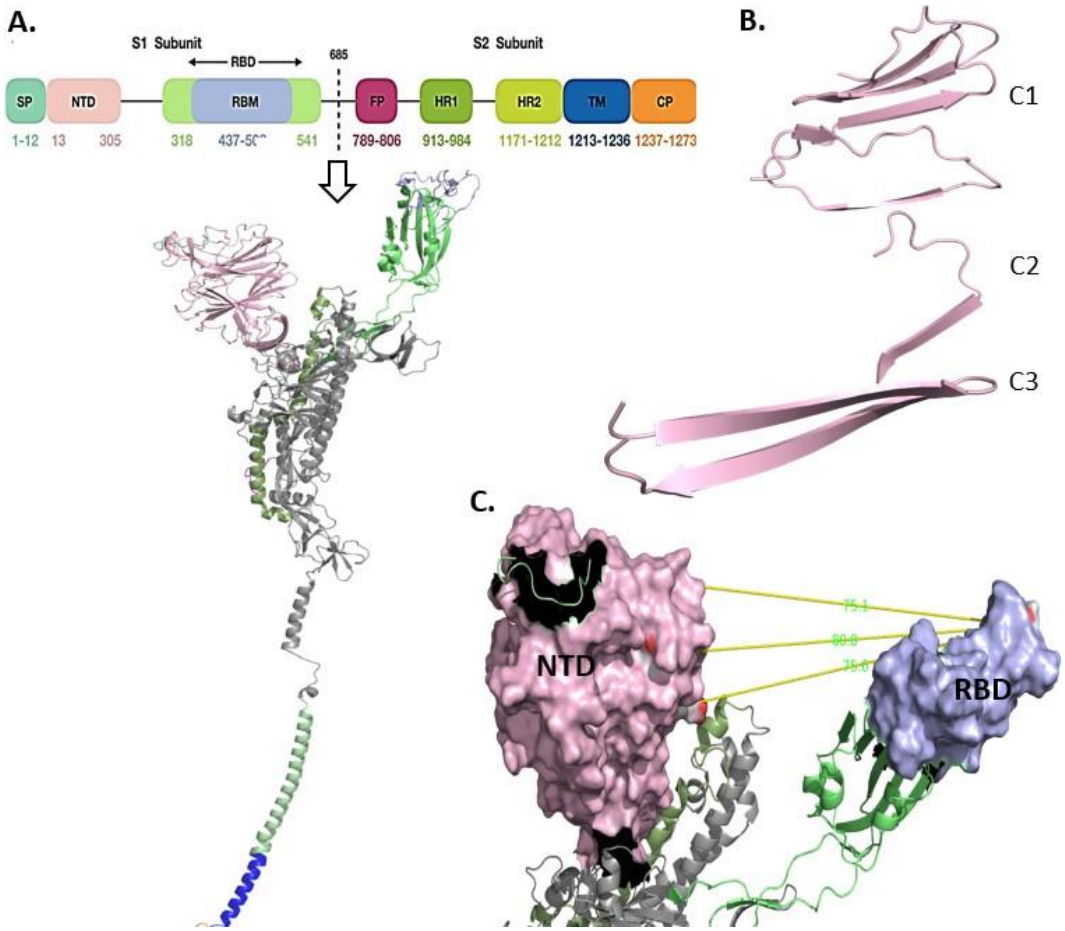


Figure 6.3 The mapping of compensating mutation regions onto the protomer of the spike(S) protein in three dimensions using RBD-up form.

(A) Different domains and subdomains belong to the S1 (upper part) and S2 (lower or base part) units of the S-protomer; (B) structural segments C1, C2, and C3 on the N-terminal domain (NTD) (pink) have been mapped and involve variously composed beta strands; and (C) a schematic presentation of the functional significance of C1, C2, and C3 that differently establish a long-range allosteric.

6.2.2 Long-range dynamic allostery is the driving force behind mutational spots in the receptor-binding domain (RBD).

We have investigated whether there are any allosteric interactions between the compensatory mutations and the various VOI-specific RBD mutations in the spike protein in order to comprehend the functional implications of the compensatory mutation network. The chemical shift changes of the diamagnetic ^{1}H, ^{13}C, and ^{15}N of protein residues can successfully be used to identify such a long-range allosteric perturbation (> 20 A^{0}), which is triggered by mutations or chemical modifications of the ligand effector and transmitted to a remote end. SHIFTX2[26] is a computer algorithm that can successfully forecast ^{1}H, ^{13}C, and ^{15}N chemical shifts from protein coordinate data (Table S3) https://pubs.acs.org/doi/full/10.1021/acsomega.1c05155. SHIFTX2 combined sequence-based and structure-based chemical shift prediction techniques to achieve high accuracy using a large, high quality database of training proteins (> 190). These advanced machine learning techniques included many more features (χ^2 and χ^3 angles, solvent accessibility, H-bond geometry, pH, and temperature). Along with the well-known NMR chemical shift covariance analysis, this chemical shift-based forecast of long-range allostery has also been investigated in Ohm-a numerically effective network-based technique[211, 212]. which resembles the popular NMR chemical shift correlation analysis (CHESCA)[205]. Based on a perturbation propagation method that repeatedly repeats the stochastic process of perturbation spreading on a network of interacting residues in a given protein, “Ohm” automatically finds the allosteric network topology and identifies allosterically coupled residues. An allosteric coupling intensity (ACI), which represents the

frequency with which each residue is impacted by a disturbance, is used to quantify this residue-residue allosteric coupling. We have computed ACI for each of the VOC/VOI-specific RBD mutation sites and compared the findings to the chemical shift-based results by designating the C1, C2, and C3 residue positions in the NTD as active sites. We have tested whether the paired inter residue correlation remains linear in various conformational states of the RBD with/without ACE2 bound and further noticed if there is any departure from this linear relationship under various RBD mutations using combined ^{1}H and ^{15}N chemical shifts. Such an inter residue allosteric coupling is indicated by the linear association in the various states of the same protein. This analysis revealed that several residues from the C1, C2, or C3 cluster of the compensating mutation network still exhibited a significant linear association with the VOC/VOI-specific RBD mutation sites in all of the specified RBD conformational states. Although the RBD mutated- and non-mutated states also exhibit this linearity, there are irregular variations between the mutated and non-mutated states in the C1, C2, and C3 regions (**Table 6.2**). When compared among the different RBD- mutated states (i.e., E484K, K417N, L452Q, L452R, N501Y, and T478K), it showed strong allosteric signals activated by suite of residues in C1, C2, and C3 with overall correlation ≥ 0.7 (**Figure 6.4** A) that are differently allosterically linked with the RBD mutation, suggesting the possibility of long-range allosteric communication between the compensatory mutation sites in the NTD and the RBD mutation site (**Figure 6.4** B).

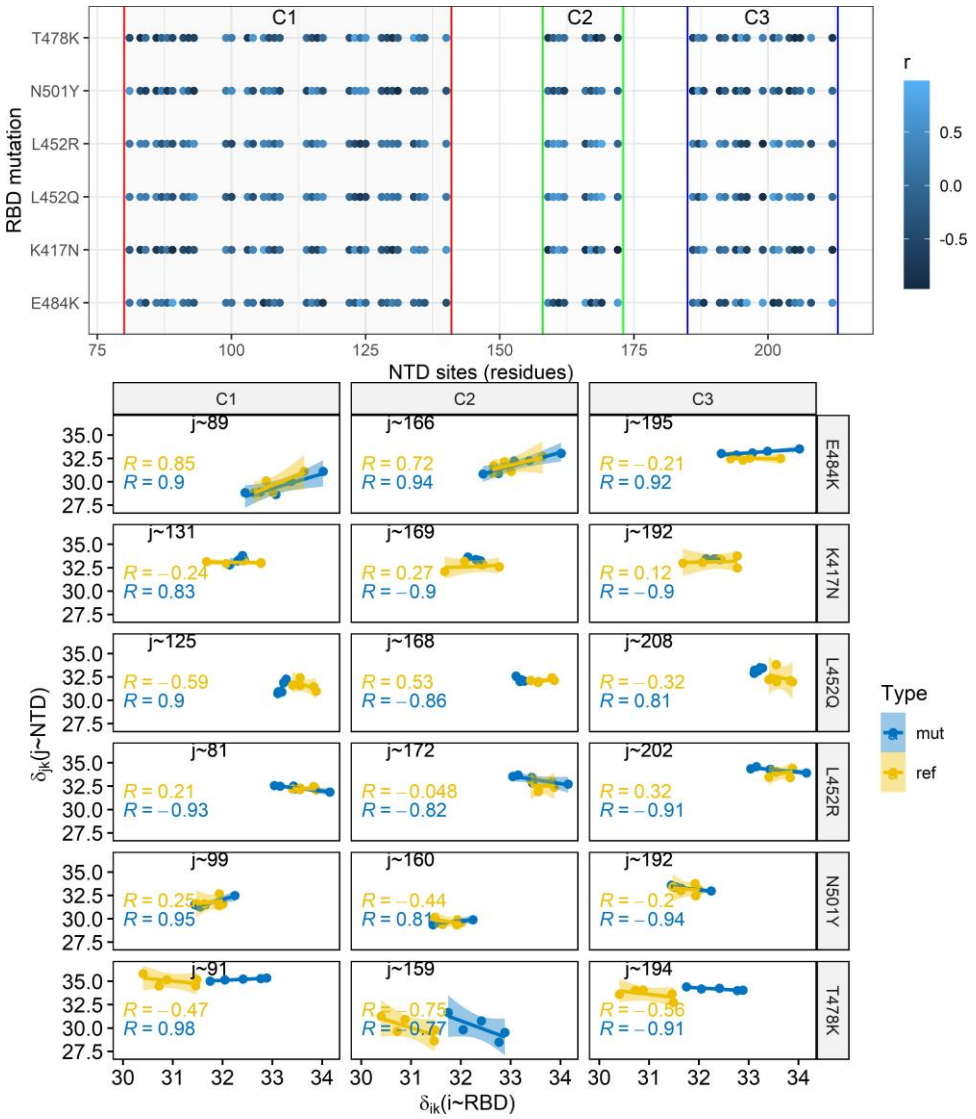


Figure 6.4 The allosteric interaction between VOI/VOC- specified RBD mutant site I and j in C1 C2 and C3

(a) Pearson association using the total chemical shift of ^{1}H and ^{15}N ppm values in the S protein's five conformational shapes, denoted by the letter "k': S-protein with various particular RBD variants, as well as RBD-down, RBD-up, bound with ACE2, and clockwise and anticlockwise motion of RBD bound with ACE2 (δ_{ik}), and (b) The best correlation sites, also known as j-residues, in the compensating mutation segments C1, C2, and C3 that exhibit an allosteric link with the particular RBD site that is affected by VOI/VOC-specific RBD mutations.

Table 6.2 The percentage of residues in the compensatory mutation regions C1,C2 and C3 with the total chemical shift-based association |r|≥0.7 and an allosteric link to the RBD site

RBD mutation	Class	Mutated con. (%)	Reference con. (%)
E484K	C1	23.53	11.76
K417N	C1	8.82	17.65
L452Q	C1	26.47	2.94
L452R	C1	17.65	2.94
N501Y	C1	20.59	11.76
T478K	C1	29.41	20.59
E484K	C2	44.44	55.56
K417N	C2	33.33	44.44
L452Q	C2	11.11	11.11
L452R	C2	22.22	11.11
N501Y	C2	11.11	33.33
T478K	C2	11.11	NA
E484K	C3	37.50	25.00
K417N	C3	18.75	18.75
L452Q	C3	18.75	12.50
L452R	C3	6.25	12.50
N501Y	C3	18.75	25.00
T478K	C3	18.75	25.00

Number of residues among the compensatory mutation clusters that allosterically coupled with the RBD mutation sites vary significantly, referring to differential effects of RBD mutation sites. For example, spike K417N and the non-mutated states involves 33.33% and 44.44% residues of the C2 cluster respectively (**Table 6.2**) that allosterically coupled with the RBD site 417 with the absolute correlation greater than 0.7. Whereas S E484K involves 44.44% residues of the C2 cluster which is 10% less than its non-mutated form, having allosteric coupling with RBD mutation site 484. Further evidence of the existence of such coupling was found in the recorded allosteric coupling intensity (ACI) in the “Ohm” webserver. We determined the

allosteric coupling intensity (ACI) for the RBD mutation sites using the inputs of the compensatory mutation sites of C1, C2, and C3 as active sites in Ohm independently (Table S4) https://pubs.acs.org/doi/full/10.1021/acsomega.1c05155. It was discovered that ACIs are greater than 0.2 with and without ACE2 bound; however, ACE2 bound reduces and the RBD mutation with ACE2 bound improves the ACI in comparison to its original RBD-up form. These Ohm results concur with a correlation study of the residues based on chemical changes. When C3, C2, and C1 are taken into account as the active sites, S K417N has the ACI 0.35, 0.31, and 0.28, respectively. This is greater than 40% when compared to its native RBD-up conformational shape, showing strong allosteric connections for the K417N S conformation (**Figure 6.5**). The K417 and N501 residues function as effector centres of allosteric interactions and hold remote sites that facilitate long-range allosteric communications in the complex, according to a recent research. This finding is in line with the findings made above[213]. Chemical shifts projection analysis was used to assess how RBD-specific alterations affected allosteric signalling and, ultimately, interactions with ACE2 (CHESPA)[208]. In CHESPA, one of the vectors A is projected onto the other vector B to quantify the shift along B, where A stands for the total residue-wise chemical shift differences between the RBD-up S and the non-mutated ACE2-bound S. The study gives the fractional shift (X) and the $\cos\theta$ value as two important residue-specific descriptors of the perturbation produced by the mutations (Methods). When S protein is attached to ACE2, the only residues with absolute $\cos\theta$ values close to 1 are appropriate sensors of allosteric activity. The RBD-mutation impacts towards an allosterically more active state are indicated by a positive fractional change X, whereas the opposite is true if X is negative. The results of CHESPA reveal that many NTD residues have positive X values with absolute cos larger than 0.9 (**Figure 6.6**),

which indicates robust allosteric activity of compensatory mutation sites in NTD and coupling with VOI-specific RBD mutation sites. All of the mutations have noticeable impacts, but K417N has revealed a large number of residues with extremely high positive X values. In addition, a significant number of mutation sites in the compensatory mutation segments C1, C2, and C3 are common and highly active ($X>1.0$ and $\cos(\theta)>0.9$) across all the RBD-mutations sites: 87,89,91,114, 115,123, 128 of C1; 159 of C2; and 187,196 of C3, indicating their crucial role in maintaining the allosteric communications. The RBD mutation sites K417, E484, and N501 correspond to a group of adaptable allosteric centres, in which small perturbations can modulate collective motions, alter the global allosteric response, and induce binding resistance, according to a recent study of dynamic profiling of binding and allosteric propensities of SARS-CoV-2 spike protein with different classes of antibodies[213].

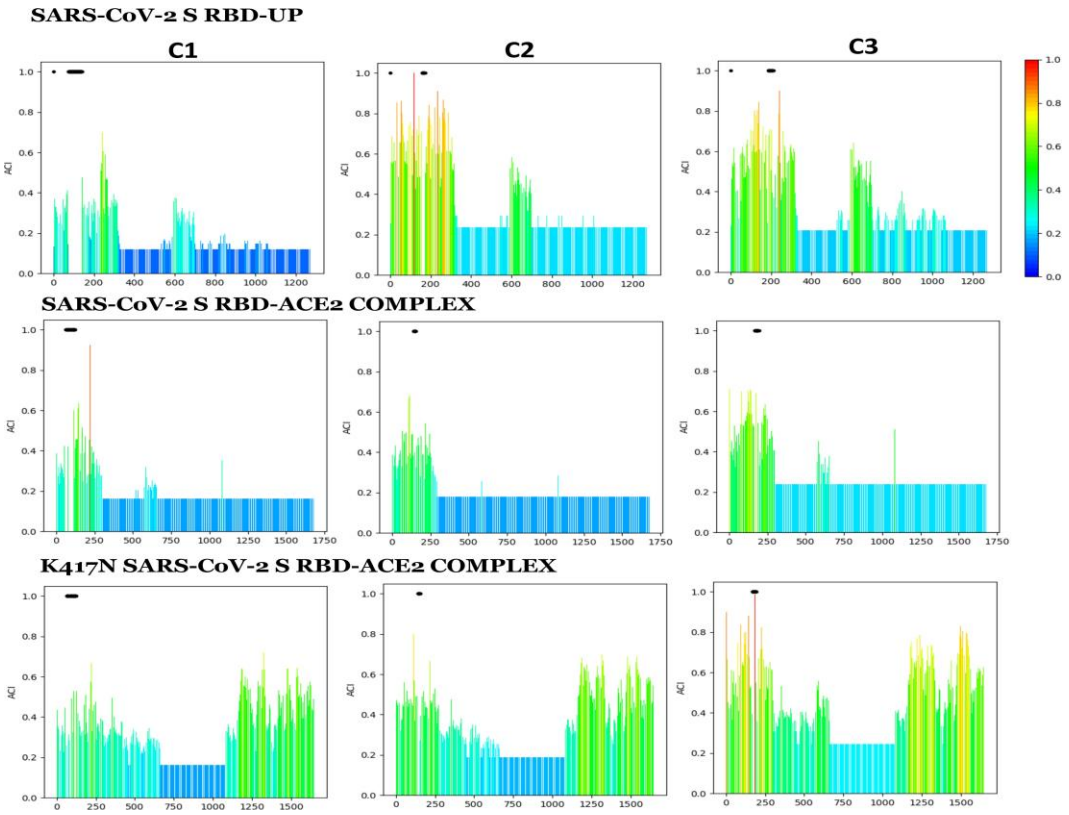
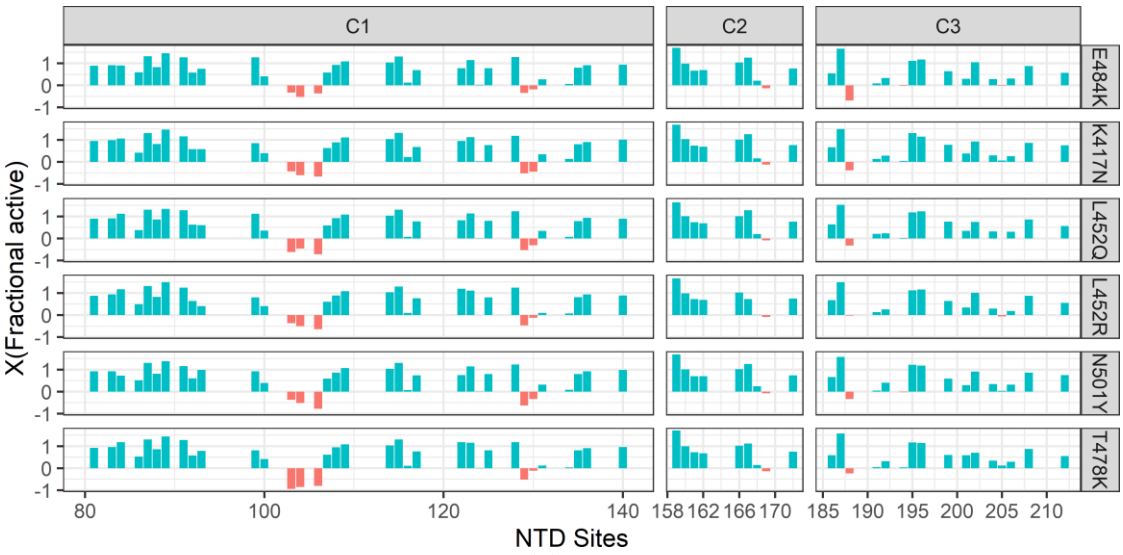
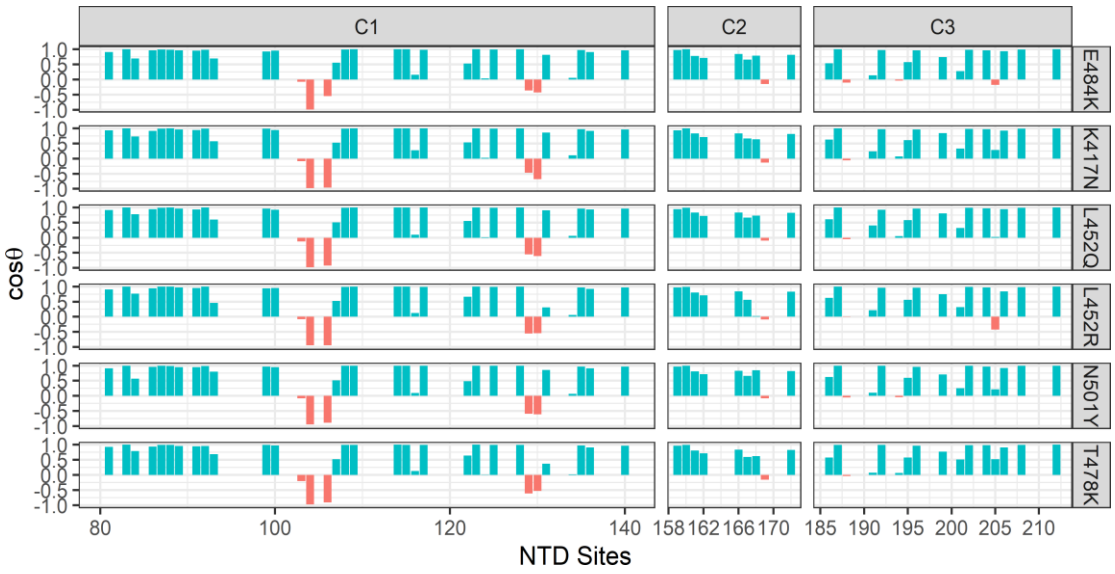


Figure 6.5 The regularity with which each residue is impacted by a change as a result of VOI/VOC- specified RBD mutations.

The residues in the compensatory mutation regions C1, C2, and C3 were used as activator sites in the "OHM webserver" to compute ACIs. It demonstrated that under the mutation K417N, the RBD mutation modifies the ACIs with a notable rise.



(a)



(b)

Figure 6.6 The chemical shift projection study demonstrating the consequences of particular RBD mutations for VOI/VOC

(a) The fractional shift (X) variance caused by the RBD mutations E484K, K417N, L452Q, L452R, N501Y, and T478K in the compensating mutation segments C1, C2, and C3. When cos θ ≥0.9, there are a few residues for which X is negative in each section, and (b) the

projection angle, $cos\theta$, which indicates whether the chemical shift is moving in a direction that favours (+ve values) or opposes (-ve values) the allosteric activity

6.2.3 Spike protein associations with ACE2 are modified by mutations through a dynamic allosteric network.

The interaction of the S-protein RBD with the human receptor ACE2 and consequent modulation of the proteolytic processing of the S1 unit for membrane fusion determines whether SARS-CoV-2 is successfully introduced into the human host cell. Here, we investigate the impact of the VOC/VOI-specific RBD mutation on this association with ACE2. In order to forecast the binding affinity from their 3D structures based on intermolecular contacts and characteristics obtained from non-interface surface, we used PROtein Binding Energy Prediction (PRODIGY)[209], a web server. Individual mutant structures are created for each of the VOI-specific RBD mutations before being entered into PRODIGY, with the standard pdb 7a94 containing one up-RBD bound to ACE2. The most abundant pose for the substitution was chosen using the Rotamer collection in UCSF-Chimera[214], and this mutated structure was then improved using the 3Drefine program[215], which maximises the hydrogen bonding network and minimises energy using all atom force fields. These energy-minimized structures were used to dock with ACE2 in the Frodock2.0 protein-docking potential docking server, selecting a dock with minimal energy and an ACE2 coupled with the up-RBD posture. Ultimately, PRODIGY used this resulting compound.

We employed PISA software to determine the molecular and chemistry characteristics of ACE2-bound RBD non-mutant- and mutant S-proteins[210].

Comparing the interface interactions reveals that all VOI-specific mutations have noticeable impacts there, with the mutated structure having a roughly 2-fold higher solvent accessible area (SAA) at the interface (**Table 6.3**), indicating a greater possibility for ACE2 and RBD interactions. Particularly, RBD mutations unique to the delta and lambda atoms (K417N, T478K, and L452Q) displayed significantly greater SAA than the active S protein bound to ACE2. The H-bond network at the junction had particular mutation-related changes, and the H-bonds were distributed unevenly. A number of salt bridges were formed at the contact in addition to the H-bond by T417N, L452Q, L452R, N501Y, and T478K. Different binding energies (ΔG) and dissociation coefficients (K_d) under various RBD mutations may be explained by this unequal distribution of H-bonding and salt-bridges. All of the RBD mutations have decreased ΔG and K_d when compared to the ACE2-bound non-mutated S-protein, suggesting greater interface interactions and binding affinity brought on by the mutations. It was discovered that K417N, out of all the changes taken into consideration, has the greatest impact on ΔG and Kd (-17.2 kcal/mol, 2.50 10-13 M). Notably, the K417N S protein maintains close contacts with ACE2 by holding 5 H-bonds with a distance cutoff of 3.00 A^0 and 4 salt bridges. Additionally, it demonstrates that K417N S, which is the biggest mutation among all mutations, is bound to ACE2 and occupies 7.7% and 3.3% of the solvent-accessible surface area at the interfaces of ACE2 and Spike, respectively. The dominant polar hydrophilic residue SER appears on the RBD 24 times out of all the interface residues, while the dominant hydrophilic residues GLU that interact with ACE2 occur 39 times, allowing for the creation of numerous H-bonds and slat bridges at the interface. We also investigated the impacts of various flexible RBD orientations, such as clockwise and anticlockwise movement of RBD, which can be caused by a number of mutations at

hinge residues close to the S1/S2 juncture, in addition to the VOC/VOI-specific mutations (e.g.,D614G)[216]. According to (**Figure 6.7** A), such variable RBD orientations have noticeable decreasing effects that change the binding affinity to ACE2. Contrary to their lowest readings for the K417N S protein, we noticed a significant rise in G and Kd under the variable RBD orientations (**Figure 6.7** B). In addition to mutations, N-glycosylation of the SARS-CoV-2 S protein is crucial for viral entrance into human cell models because viruses without N-glycans penetrate the host cell less efficiently. In a recent research, the glycosylation profile and alterations that occurred throughout the worldwide transmission of the SARS-CoV were described and contrasted. It listed 3 O-glycosylation sites that are specific to SARS-CoV-2 and 9 anticipated N-glycosylation sites, but there hasn't been any evidence of glycan site diversity thus far[217]. Further investigation revealed that the existence of glycans at sites N165 and N234 influences the RBD's conformational flexibility because they maintain the RBD's "up" shape, enabling effective binding to the human angiotensin-converting enzyme 2 (hACE2) receptor. A conformational change of the RBD towards the "down" state, which weakens accessibility to ACE2, caused by the deletion of these glycan residues through the N165A and N234A alterations, resulted in a substantial reduction in S protein binding to ACE2.

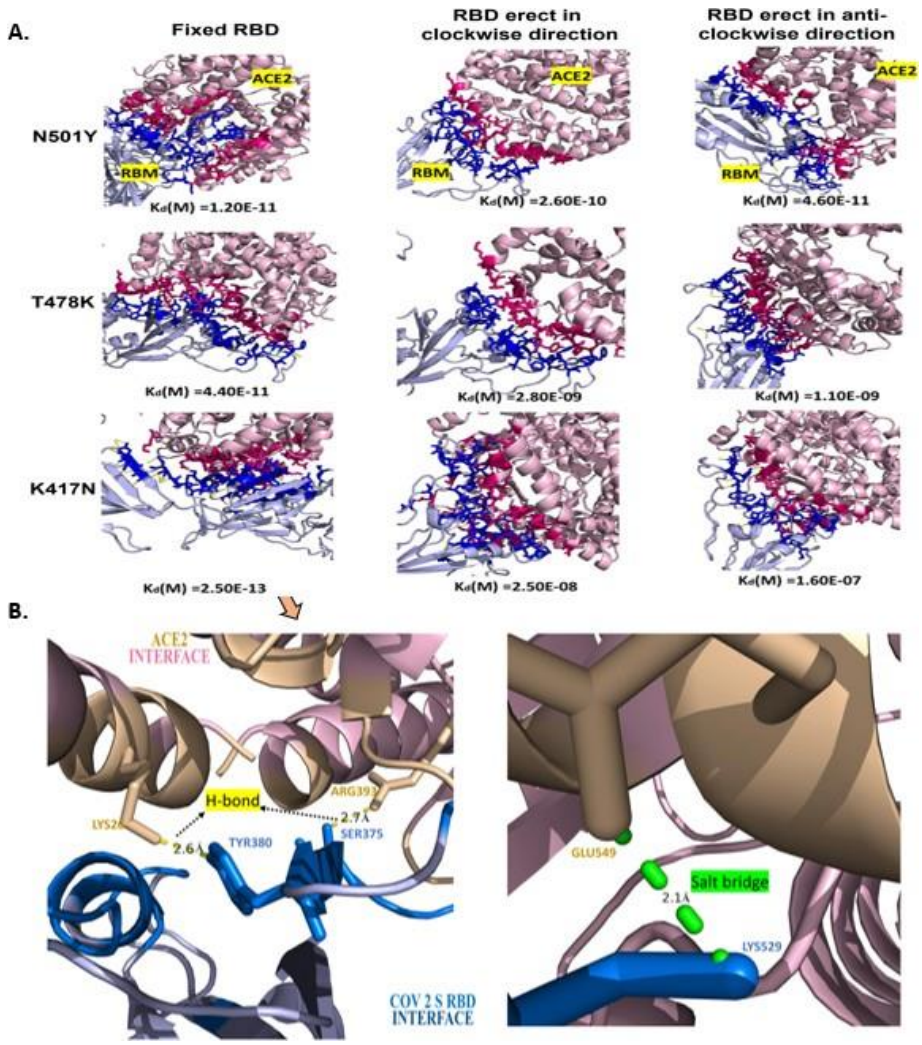


Figure 6.7 Impact of SARS-CoV-2 RBD and ACE2 interactions with various RBD angles and VOI/VOC-specified RBD mutations.

(A) Screenshots of the RBD mutations N501Y, T478K, and K417N; interface interactions and its spread between ACE2 (red) and the RBD (blue) varied among the mutated states with distinct RBD motions, resulting in various dissociation constant K_d; The K417N mutation had the lowest K_d of the six RBD mutations that were taken into consideration, as seen in (B) zoomed images of the K417N mutation's contacts with H-bonds and Salt bridges. It has the most surface contact residues (182) of the six variants and holds 5 H-bonds with a distance cut-off of 3.00 A^0. It also holds 4 salt bridges.

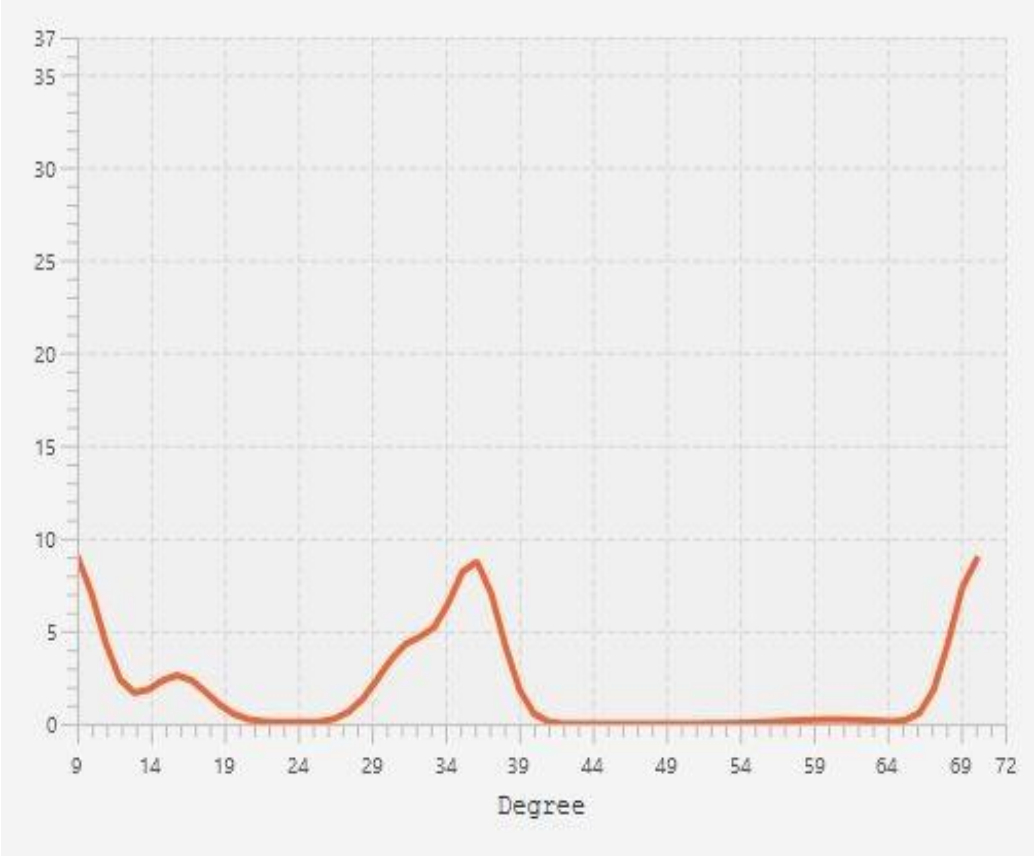


Figure 6.8 Multimodal degree distribution

.

Table 6.3 Interface characteristics of the S-RBD and ACE2 interactions under Six RBD variants with various RBD orientations that are unique to VOI/VOC

	Interface Interaction Properties		Wild	E484K	K417N	L452Q	L452R	N501Y	T478K
	K_d (M) at 25^{0}C		2.50E-09	1.90E-11	2.50E-13	1.90E-10	4.60E-11	1.20E-11	4.40E-11
	Δ G (kcal/mol)		-11.7	-14.6	-17.2	-13.3	-14.1	-14.9	-14.1
	SAA at interface (%)	ACE2	3.1	4.7	7.7	6.3	5.1	4.9	5.2
		SARS-CoV-2 S	1.4	2.1	3.6	2.8	2.4	2.4	2.4
ACE2-bound stable RBD	No. of intermolecular contacts		72	105	182	162	109	124	111
	No. H-bonds (dist. cut-off: 3 Å)		6	2	5	5	2	5	2
	No. of salt bridges		0	0	4	5	1	4	1
ACE2 bound-RBD moved clockwise	K_d		3.30E-09	6.70E-09	2.50E-08	1.20E-09	1.70E-09	2.60E-10	2.80E-09
	Δ G		-11.6	-11.1	-10.4	-12.1	-11.9	-13.1	-11.7
ACE2 bound-RBD moved anticlockwise	K_d		3.50E-09	2.10E-08	1.60E-07	3.30E-08	1.60E-08	4.60E-11	1.10E-09
	Δ G		-11.5	-10.5	-9.3	-10.2	-10.6	-14.1	-12.2

6.3 Methods

6.3.1 Detecting mutational sites that are susceptible across SARS-CoV-2 spike types

To find susceptible mutation sites that are subject to frequent changes in the known SARS-CoV-2 variants, we depend on differential sequence comparison among the mutated sequences rather than traditional reference-based sequence analysis. With the idea that SARS-CoV-2 would change physically for a tighter bond with the host, we only focused on the spots that are extremely prone to mutation. For the convenience of computation for such extremely susceptible sites, we used a binary vector-based comparison. For each amino acid, two alleles are examined. A match receives a score of 0 and is removed from the final vector. Mismatch, on the other hand, is awarded with 1 to show that there is alternation in a specific location. The same procedure was used to create N binary vectors from every combination of $\binom{N}{2}$ potential sequences.

Next, we determine each site's probability of exposure.

Suppose that there are M = $\binom{N}{2}$ binary vectors, each of which has L leftover sites. In this case, S = {s_1, s_2,…,s_L}; where s_i ={ b_1, b_2,…, b_M} and so on. The following formula can be used to determine how vulnerable a mutant site s_i is.

$$P(s_i) = \frac{\sum_1^M b_i}{M} \qquad 6.1$$

If P (s_i) >τ, a user-defined cutoff, a site si is regarded as possibly susceptible to mutation. By taking into account all of these possibly susceptible sites, we create a network of vulnerable sites that coexist.

6.3.2 The extraction of cohesive mutant sites.

In relation to all potential variants, we determined concurrent vulnerable sites that show specific cooccurrence of alternation. To determine coexisting susceptible locations, we used the conditional probability score. The following formula can be used to calculate the probability that two locations, si and sj (s_i, …, s_j), will cohabit.

s_i, …, s_j if $P(s_i|s_j)= \theta$ *or* $P(s_j|s_i)= \theta$ $(0 \leq \theta \leq 1)$,

$$\mathbf{P(si|sj)} = \frac{\mathbf{P}(\boldsymbol{s_i}\mathbf{and}\ \boldsymbol{s_j})}{\mathbf{P}(\boldsymbol{s_j})} \qquad 6.2$$

As shown below, an adjacency matrix A is created by combining all pairs of s_i, ..., s_j.

$$A(i,j) = \begin{cases} 1, & if\ s_i, \dots, s_j\ exists \\ 0, & otherwise \end{cases}$$

After that, we used Cytoscape plugins[206] and the well-known molecular complex detection (MCODE) technique[207] to select highly linked subgraphs of vulnerable sites from the aforementioned network. In fact, MCODE is made to find interconnected areas in sizable protein–protein interaction networks that could be molecular complexes. The technique relies on vertex weighting by local neighbourhood density and uses an outer traversal from a seed protein that is locally dense to separate the dense areas according to predetermined parameters.

6.3.3 Residues between RBD and NTD undergo chemical shift changes and allosteric coupling

Determining a network of residues that mediates the cross-talk between distant locations is still frequently a difficult experimental task. As in this case, the allosteric signal propagation depends on subtle but crucial conformational and side-chain packing rearrangements that frequently fall below the resolution of conventional X-ray or NMR structure determination methods, such clusters of coupled residues are particularly difficult to identify. It was mentioned that chemical shift (CS) data can be investigated to look at long-range allosteric interactions in order to deal with this scenario. It was discovered that CS data exploration was very successful because allosterically linked residues show coordinated and correlated chemical shift changes for a given collection of disturbances. Such correlations can be seen in a group of allosteric changes that, while requiring only minor covalent modifications, cause varying levels of activation and are spatially colocalized within a single area of the protein structure. For a specific collection of disturbances, allosterically linked residues show coordinated and correlated chemical shift changes. These associations can be seen in a collection of allosteric perturbations that colocalize physically within a single area of the protein structure and cause various levels of activation with only minor covalent modifications. Chemical changes for various residues suitably far from the effector binding site 56 that detects the same perturbed equilibrium are linearly linked in a two-state activation model with a fast exchange regime. Because the spots that correlate to the same active ligands were partially rearranged in the multistate model, linearity was still preserved. These assumptions allow for the effective implementation of chemical shift covariance analysis, which allows for the

probing of the presence of long-range allosteric communication by observing the linear coupling of distant residues by their chemical shifts (CHESCA)[205]. Long-range allosteric signal transmission depends critically on both structural modifications and effector-dependent dynamics modulations. The terminal receptor sites of these allosteric signals can be successfully identified using comparative studies of the structural and dynamic characteristics of apo and effector-bound states. Technical challenges still exist in delineating a network of residues that mediates this cross-talk between distant locations. Finding such groups of coupled residues is particularly difficult when the allosteric signal propagation depends mainly on conformational and side-chain packing rearrangements. Chemical shifts have been shown to be very efficient for determining long-range allostery larger than 20 angstrom because they are extremely sensitive to both structural changes and the effector-dependent modulations. Residues that are part of the same effector-dependent allosteric network show a coordinated reaction to the perturbation set when using a chemical-shift-based method, whereas this may not be the case when the network was simply identified based on the 3D protein structure. We separately examined whether there was a linear correlation between the residues in segments C1, C2, and C3 of the compensatory mutation network in the NTD and the VOC-/VOI-specific RBD mutation sites away from the ACE2- bound residues given the five conformational states of the spike protein with active/inactive RBD, RBD bound with ACE2, and clockwise and anticlockwise poses of ACE2-bound RBD.

Chemical shifts that are computed for each residue individually are calculated as the weighted total of the amide proton ^{1}H and ^{15}N nitrogen ppm values.

$$\delta_{ik} = W^N \delta_{ik}^N + W^H \delta_{ik}^H \quad 6.3$$

where W^N and W^H are the weights for the chemical shifts δ_{ik}^N and δ_{ik}^H, respectively, and δ_{ik}^N indicates the total chemical shift of residue i at the k^{th} perturbed state. The perturbation-dependent chemical shift variations of two residues (or sites) i and j show a linear correlation if they are members of the same allosteric network, independent of their size. As a result, residues i and j that are allosterically linked create the following linear equation.

$$\delta_{ik} = \delta_{jk}\alpha + \beta \quad 6.4$$

Based on the finding that a small RBD reorientation causes correlated perturbation in the immediate environment of residues i and j, nonlinear components in eq. 6.4 are disregarded. The maintained correlation $|r_{ij}| \geq 0.7$ illustrates such correlated disturbances (Pearson correlation (r_{ij})) as computed in eq. 6.5, showing a coordinated group reaction to perturbed states.

$$r_{ij} = \frac{\sum(\delta_{ik} - \overline{\delta_{ık}})(\delta_{jk} - \overline{\delta_{jk}})}{\sqrt{\sum(\delta_{ik} - \overline{\delta_{ık}})^2(\delta_{jk} - \overline{\delta_{jk}})^2}} \quad 6.5$$

where δ_{ik} and δ_{jk} are two vectors of identical length and $\overline{\delta_{ık}}$ and $\overline{\delta_{jk}}$ are the corresponding means of δ_{ik} and δ_{jk}.

We used chemical shift projection analysis to determine how the particular RBD mutation impacts the allosteric active states (CHESPA)[208]. Commonly used compounded ppm changes, calculated as $\sqrt{(0.2\Delta\delta N)^2 + (\Delta\delta H)^2}$,[218-223], only take

into account the size of chemical shift differences brought on by a mutation, not how the mutation specifically impacts the dynamic equilibrium. The magnitude of vector A, which connects the apo-S and S-mutant ACE2-bound peaks and is specified in the plane of the ^{1}H and scaled ^{15}N ppm coordinates, was used to determine the cumulative chemical shift difference between the apo-S and the S-mutant. The ^{15}N ppm readings have a scaling factor of 0.2. The magnitude of vector B, the activation vector connecting the apo/inactive to the allosterically active state, is used to calculate the compounded chemical-shift differential between the apo-S and the ACE2-bound S. The change along the activation vector brought on by a specific mutation is measured by the projection of vector A onto vector B. The fractional shift (X), which is computed as the ratio of the component of vector A along vector B and the amplitude of vector B (i.e., |B|), is used to determine the degree of activation (or inactivation) brought about by a mutation. The cos θ number is a counterpart to the scalar fractional shift (X). It is founded on the angular relationship between vectors A and B. As a consequence, two important residue-specific markers of the perturbation induced by the mutation—the fractional shift (X) and the cos θ are produced by the projection analysis of the chemical shifts. The fractional shift, X, is either positive or negative depending on whether the mutation moves the balance towards the allosterically active state. When the mutation causes ppm changes of similar size and direction, |X|~ 1, the absolute value of X approaches 0 if the ppm variations produced by the mutation are minimal. With the powerful allosteric impact of the mutation, the |cos θ|

values get closer to unity (i.e., |cos θ| ~1). As opposed to long-range allostery, |cos θ |< 1 for residues is more substantially impacted by the mutation through nearest-neighbour effects or other structural changes brought on by the mutation.

The ratio of the component of vector A along vector B to the magnitude of vector B, or |B|, is used to compute the fractional shift (X)[208],

$$X = \frac{|\vec{A}|cos\theta}{|\vec{B}|} \qquad 6.6$$

Where, A=[$0.2\Delta\delta_{ik_r}{}^{N}, \Delta\delta_{ik\prime_m}{}^{H}$] , B=[$0.2\Delta\delta_{ik_r}{}^{N}, \Delta\delta_{ik\prime_r}{}^{H}$] and m and r denotes mutated and reference conditions, respectively, in kth (or k'th) state. θ measures angle between Vector A and B,

$$cos\theta = \frac{\vec{A}.\vec{B}}{|\vec{A}|.|\vec{B}|} \qquad 6.7$$

$$X = \frac{\vec{A}.\vec{B}}{|\vec{B}|^2} \qquad 6.8$$

6.3.4 Utilizing 3D protein structures for interface analysis

Using the PROtein binDIng enerGY Prediction (PRODIGY)[209] and PDBePISA[210] online services, RBD and ACE2 binding affinity, dissociation constant, H-bond and salt-bridge calculations were performed. With the quantity and type of intermolecular contacts within the 5.5 A^0 distance limit, PRODIGYe forecasts binding affinity and identifies the interfaces from 3D protein structures. Based on a straightforward linear regression of interface contacts (ICs) and a few characteristics

of non-interacting surfaces (NIS), which have been shown to affect the binding affinity.

$$\Delta G_{predicted} = -0.09459 \times ICs_{charged/charged}$$

$$- 0.10007 \times ICs_{charged/apolar}$$

$$+ 0.19577 \times ICs_{polar/polar} - 0.22671$$

$$\times ICs_{polar/apolar} + 0.18681 \times \%$$

$$NIS_{apolar} + 0.3810 \times \%NIS_{charged} - 15.9433.$$

The sort of contacts within a radius of 5.5 A^0 are used to categorise ICs, and the number of interfacial contacts (ICs) discovered at the interface between Interactor1 and Interactor2 is represented by the ratio $ICs_{xxx/yyy}$. Using the equation: $\Delta G = RT \ln K_d$, where R is the ideal gas constant (in kcal K^{-1} mol^{-1}), T is the temperature (in K), and ΔG is the expected free energy, one can determine the dissociation constant (K_d). The setting for the temperature in our estimate is 25.0 °C. We used the PDBePISA website to derive interactions between salt bridges and H-bonds at the interface. If the spacing between the heavy elements in the donor and acceptor is less than 3.89 A^0, PISA takes this into account when determining the presence of a H bond. For a salt bridge, the pertinent distance is 4 A^0. In FRODOCK webserver[224], ACE2 docking with the RBD-mutated spike was performed. It employs a fast-rotational binding technique using the 3D coordinates of two cooperating proteins. It employs a quick circular docking technique using the three-dimensional coordinates of two binding proteins. It carried out a six-dimensional (6D)

rigid-body comprehensive search of the orientations of a stationary molecule about a mobile receptor (3D rotations + 3D translations) in order to optimise global energy. Each energy term was computed using a correlation function specified by the interplay of potential receptor and ligand components, only taking into account the rotational component. An implied translational scan was used in conjunction with this quick comprehensive rotational search. By evenly sampling the space with a matrix of set step sizes, the translational search was performed tacitly.

6.4 Summary

- The SARS-CoV-2 spike protein has several variant-specific mutations, including E484K, K417N, L452Q, L452R, N501Y, and T478K, that affect its structure and function.

- These mutations are not random, but are part of an allosteric network that affects interactions between the spike protein and human receptor ACE2, leading to higher transmissibility and infectivity.

- Compensatory mutations in the N-terminal domain (NTD) are also involved in this network, and are allosterically coupled with specific RBD-mutation sites.

- Mutations in the RBD increase interactions with ACE2 to varying extents, depending on their allosteric connections with compensatory mutation clusters in the NTD.

- K417N has the largest effect on allostery and the highest binding affinity with ACE2, which may explain why the delta variant is highly transmissible.

- Understanding the significance of these mutations can aid in targeted control measures, laboratory characterization, and therapeutic efforts.

Future direction

Future aim is to explore highly heterogeneous and high throughput molecular data and apply this knowledge in the context of cancer, autoimmune diseases, and other areas.

Here, I want to learn more about the molecular processes at play in complicated illnesses like cancer. Utilizing cutting-edge technologies, this strategy generates significant quantities of molecular data from a variety of sources, including genomics, transcriptomics, proteomics, and metabolomics. Another area of focus would be Autoimmune disorders, which are defined by an abnormal immune reaction that targets healthy cells in the body, are a different field of emphasis. We intend to find important molecular mechanisms that add to the onset and progression of illness by examining molecular data from individuals with autoimmune disorders.

Further, I would also like to extend my research using different machine ML and AI tools.

When examining molecular data, machine learning (ML) and artificial intelligence (AI) tools can be very useful because they can spot trends and connections that might not be apparent to people.

Bibliography

1. Tan, A.C. and D. Gilbert, *Ensemble machine learning on gene expression data for cancer classification.* 2003.

2. Kulkarni, M.M., *Digital multiplexed gene expression analysis using the NanoString nCounter system.* Current protocols in molecular biology, 2011. **94**(1): p. 25B. 10.1-25B. 10.17.

3. Castano, S. and V. De Antonellis, *Global viewing of heterogeneous data sources.* IEEE Transactions on Knowledge and Data Engineering, 2001. **13**(2): p. 277-297.

4. Li, H., F.-l. Chung, and S. Wang, *A SVM based classification method for homogeneous data.* Applied Soft Computing, 2015. **36**: p. 228-235.

5. Yang, X., et al., *Novel financial capital flow forecast framework using time series theory and deep learning: a case study analysis of Yu'e Bao transaction data.* Ieee Access, 2019. **7**: p. 70662-70672.

6. Shekhar, S., et al., *Spatiotemporal data mining: A computational perspective.* ISPRS International Journal of Geo-Information, 2015. **4**(4): p. 2306-2338.

7. Elliot, P., et al., *Spatial epidemiology: methods and applications*. 2000: Oxford University Press.

8. Sanchez-Lengeling, B. and A. Aspuru-Guzik, *Inverse molecular design using machine learning: Generative models for matter engineering.* Science, 2018. **361**(6400): p. 360-365.

9. Paz Ocaranza, M., et al., *Counter-regulatory renin–angiotensin system in cardiovascular disease.* Nature Reviews Cardiology, 2020. **17**(2): p. 116-129.

10. Fattah, C., et al., *Gene therapy with angiotensin-(1-9) preserves left ventricular systolic function after myocardial infarction.* Journal of the American College of Cardiology, 2016. **68**(24): p. 2652-2666.

11. Ocaranza, M.P., et al., *Angiotensin-(1–9) reverses experimental hypertension and cardiovascular damage by inhibition of the angiotensin converting enzyme/Ang II axis.* Journal of hypertension, 2014. **32**(4): p. 771-783.

12. Santos, R.A., et al., *Angiotensin-(1–7) is an endogenous ligand for the G protein-coupled receptor Mas.* Proceedings of the National Academy of Sciences, 2003. **100**(14): p. 8258-8263.

13. Schinzari, F., et al., *Favorable vascular actions of angiotensin-(1–7) in human obesity*. Hypertension, 2018. **71**(1): p. 185-191.

14. Carey, R.M., *Cardiovascular and renal regulation by the angiotensin type 2 receptor: the AT2 receptor comes of age.* Hypertension, 2005. **45**(5): p. 840-844.

15. Kemp, B.A., et al., *Renal AT2 receptors mediate natriuresis via protein phosphatase PP2A.* Circulation Research, 2022. **130**(1): p. 96-111.

16. van Esch, J.H., et al., *AT2 receptor-mediated vasodilation in the mouse heart depends on AT1A receptor activation.* British journal of pharmacology, 2006. **148**(4): p. 452-458.

17. Ali, Q., S. Patel, and T. Hussain, *Angiotensin AT2 receptor agonist prevents salt-sensitive hypertension in obese Zucker rats.* American Journal of Physiology-Renal Physiology, 2015. **308**(12): p. F1379-F1385.

18. Ferrario, C.M., *Role of angiotensin II in cardiovascular disease - Therapeutic implications of more than a century of research.* Journal of the Renin-Angiotensin-Aldosterone System, 2006. **7**(1): p. 3-14.

19. Atlas, S.A., *The renin-angiotensin aldosterone system: pathophysiological role and pharmacologic inhibition.* J Manag Care Pharm, 2007. **13**(8 Suppl B): p. 9-20.

20. Ranjit, A., S. Khajehpour, and A. Aghazadeh-Habashi, *Update on Angiotensin II Subtype 2 Receptor: Focus on Peptide and Nonpeptide Agonists.* Mol Pharmacol, 2021. **99**(6): p. 469-487.

21. Carey, R.M., Z.-Q. Wang, and H.M. Siragy, *Novel actions of angiotensin II via its renal type-2 (AT2) receptor.* Current Hypertension Reports, 1999. **1**(2): p. 151-157.

22. Miyata, N., et al., *Distribution of angiotensin AT1 and AT2 receptor subtypes in the rat kidney.* Am J Physiol, 1999. **277**(3): p. F437-46.

23. Ozono, R., et al., *Expression of the subtype 2 angiotensin (AT2) receptor protein in rat kidney.* Hypertension, 1997. **30**(5): p. 1238-46.

24. Padia, S.H. and R.M. Carey, *AT2 receptors: beneficial counter-regulatory role in cardiovascular and renal function.* Pflügers Archiv-European Journal of Physiology, 2013. **465**(1): p. 99-110.

25. Carey, R.M., Z.Q. Wang, and H.M. Siragy, *Role of the angiotensin type 2 receptor in the regulation of blood pressure and renal function.* Hypertension, 2000. **35**(1 Pt 2): p. 155-63.

26. Siragy, H.M., et al., *Angiotensin subtype-2 receptors inhibit renin biosynthesis and angiotensin II formation.* Hypertension, 2005. **45**(1): p. 133-7.

27. Carey, R.M., *AT2 receptors: potential therapeutic targets for hypertension.* American journal of hypertension, 2017. **30**(4): p. 339-347.

28. McCarthy, C.A., et al., *Direct angiotensin AT2 receptor stimulation using a novel AT2 receptor agonist, compound 21, evokes neuroprotection in conscious hypertensive rats.* PloS one, 2014. **9**(4): p. e95762.

29. Siddiqui, A.H., Q. Ali, and T. Hussain, *Protective role of angiotensin II subtype 2 receptor in blood pressure increase in obese Zucker rats.* Hypertension, 2009. **53**(2): p. 256-261.

30. Higuchi, S., et al., *Angiotensin II signal transduction through the AT1 receptor: novel insights into mechanisms and pathophysiology.* Clinical science, 2007. **112**(8): p. 417-428.

31. Mehta, P.K. and K.K. Griendling, *Angiotensin II cell signaling: physiological and pathological effects in the cardiovascular system.* American Journal of Physiology-Cell Physiology, 2007. **292**(1): p. C82-C97.

32. Nguyen Dinh Cat, A. and R.M. Touyz, *Cell signaling of angiotensin II on vascular tone: novel mechanisms.* Current hypertension reports, 2011. **13**(2): p. 122-128.

33. Funke-Kaiser, H., et al., *Adapter proteins and promoter regulation of the angiotensin AT2 receptor—implications for cardiac pathophysiology.* Journal of the Renin-Angiotensin-Aldosterone System, 2010. **11**(1): p. 7-17.

34. Pilvankar, M.R., M.A. Higgins, and A.N. Ford Versypt, *Mathematical model for glucose dependence of the local renin–angiotensin system in podocytes.* Bulletin of mathematical biology, 2018. **80**(4): p. 880-905.

35. Pilvankar, M.R., H.L. Yong, and A.N. Ford Versypt, *A glucose-dependent pharmacokinetic/pharmacodynamic model of ace inhibition in kidney cells.* Processes, 2019. **7**(3): p. 131.

36. Pérez-Rosas, N. and J. Rodriguez-González. *Pharmacological Modulation of the Renin-Angiotensin System by Mathematical Modeling*. in *Proceedings of the Western Pharmacology Society*. 2011.

37. Versypt, A.N.F., G.K. Harrell, and A.N. McPeak, *A pharmacokinetic/pharmacodynamic model of ACE inhibition of the renin-angiotensin system for normal and impaired renal function.* Computers & Chemical Engineering, 2017. **104**: p. 311-322.

38. Lo, A., et al., *Using a systems biology approach to explore hypotheses underlying clinical diversity of the renin angiotensin system and the response to antihypertensive therapies*, in *Clinical trial simulations*. 2011, Springer. p. 457-482.

39. Pérez-Rosas, N. and J. Rodríguez-González. *Pharmacological Modulation of the Renin-Angiotensin System by Mathematical Modeling*. in *Proc. West. Pharmacol. Soc*. 2011.

40. Behera, B.K., et al., *Polycyclic Aromatic Hydrocarbons (PAHs) in inland aquatic ecosystems: Perils and remedies through biosensors and bioremediation.* Environmental pollution, 2018. **241**: p. 212-233.

41. Matos, J., C. Silveira, and M. Cerqueira, *Particle-bound polycyclic aromatic hydrocarbons in a rural background atmosphere of southwestern Europe.* Science of The Total Environment, 2021. **787**: p. 147666.

42. Ofori, S.A., S.J. Cobbina, and D.A. Doke, *The occurrence and levels of polycyclic aromatic hydrocarbons (PAHs) in African environments—a systematic review.* Environmental Science and Pollution Research, 2020. **27**: p. 32389-32431.

43. Cao, C., et al., *Prenatal exposure to polycyclic aromatic hydrocarbons could increase the risk of low birth weight by affecting the DNA methylation states in a Chinese cohort.* Reproductive Biology, 2021. **21**(4): p. 100574.

44. Gan, N., L. Martin, and W. Xu, *Impact of polycyclic aromatic hydrocarbon accumulation on oyster health.* Frontiers in Physiology, 2021. **12**: p. 734463.

45. Khanverdiluo, S., E. Talebi-Ghane, and A. Heshmati, *The concentration of polycyclic aromatic hydrocarbons (PAHs) in mother milk: a global systematic review, meta-analysis and health risk assessment of infants.* Saudi journal of biological sciences, 2021. **28**(12): p. 6869-6875.

46. Zhang, R., et al., *Occurrence, distribution, and fate of polychlorinated biphenyls (PCBs) in multiple coral reef regions from the South China Sea: A case study in spring-summer.* Science of The Total Environment, 2021. **777**: p. 146106.

47. Alegbeleye, O.O., B.O. Opeolu, and V.A. Jackson, *Polycyclic aromatic hydrocarbons: a critical review of environmental occurrence and bioremediation.* Environmental management, 2017. **60**: p. 758-783.

48. Parinos, C., et al., *Sources and downward fluxes of polycyclic aromatic hydrocarbons in the open southwestern Black Sea.* Organic geochemistry, 2013. **57**: p. 65-75.

49. Wang, Y., et al., *Source apportionment of polycyclic aromatic hydrocarbons (PAHs) in the air of Dalian, China: Correlations with six criteria air pollutants and meteorological conditions.* Chemosphere, 2019. **216**: p. 516-523.

50. Nasher, E., et al., *Concentrations and sources of polycyclic aromatic hydrocarbons in the seawater around Langkawi Island, Malaysia.* Journal of Chemistry, 2013. **2013**.

51. Abayi, J.J.M., et al., *Polycyclic aromatic hydrocarbons in sediments and fish species from the White Nile, East Africa: Bioaccumulation potential, source apportionment, ecological and health risk assessment.* Environmental Pollution, 2021. **278**: p. 116855.

52. Honda, M. and N. Suzuki, *Toxicities of polycyclic aromatic hydrocarbons for aquatic animals.* International Journal of Environmental Research and Public Health, 2020. **17**(4): p. 1363.

53. Liu, L.-Y., et al., *Anthropogenic activities have contributed moderately to increased inputs of organic materials in marginal seas off China.* Environmental science & technology, 2013. **47**(20): p. 11414-11422.

54. Wu, Z., et al., *Sedimentary records of polychlorinated biphenyls in the East China Marginal Seas and Great Lakes: Significance of recent rise of emissions in China and environmental implications.* Environmental Pollution, 2019. **254**: p. 112972.

55. Feng, J., et al., *Impact of suspended sediment on the behavior of polycyclic aromatic hydrocarbons in the Yellow River: spatial distribution, transport and fate.* Applied Geochemistry, 2018. **98**: p. 278-285.

56. Wang, Z., et al., *Urban fractionation of polycyclic aromatic hydrocarbons from Dalian soils.* Environmental chemistry letters, 2012. **10**: p. 183-187.

57. Sun, Y.-X., et al., *Halogenated organic pollutants in marine biota from the Xuande Atoll, South China Sea: Levels, biomagnification and dietary exposure.* Marine pollution bulletin, 2017. **118**(1-2): p. 413-419.

58. Wang, P., Q. Li, and C.-F. Li, *General Outline of the China Seas*, in *Developments in Marine Geology*. 2014, Elsevier. p. 11-72.

59. Thomas, G.E., et al., *Evaluation of polycyclic aromatic hydrocarbon pollution from the HMS royal oak shipwreck and effects on sediment microbial community structure.* Frontiers in Marine Science, 2021. **8**: p. 650139.

60. Cunderlik, J.M. and D.H. Burn, *Switching the pooling similarity distances: Mahalanobis for Euclidean.* Water Resources Research, 2006. **42**(3).

61. R Core Team, R., *R: A language and environment for statistical computing.* 2013.

62. Salata, S. and C. Grillenzoni, *A spatial evaluation of multifunctional Ecosystem Service networks using Principal Component Analysis: A case of study in Turin, Italy.* Ecological Indicators, 2021. **127**: p. 107758.

63. Xu, Z., et al., *Spatiotemporal heterogeneity of antibiotic pollution and ecological risk assessment in Taihu Lake Basin, China.* Science of the total environment, 2018. **643**: p. 12-20.

64. Oksanen, J., et al., *The vegan package.* Community ecology package, 2007. **10**(631-637): p. 719.

65. Azid, A., et al., *Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia.* Water, Air, & Soil Pollution, 2014. **225**: p. 1-14.

66. Colomer-Vidal, P., et al., *Plant uptake of perfluoroalkyl substances in freshwater environments (Dongzhulong and Xiaoqing Rivers, China).* Journal of Hazardous Materials, 2022. **421**: p. 126768.

67. Kalf, D.F., T. Crommentuijn, and E.J. van de Plassche, *Environmental quality objectives for 10 polycyclic aromatic hydrocarbons (PAHs).* Ecotoxicology and environmental safety, 1997. **36**(1): p. 89-97.

68. Cao, Z., et al., *Distribution and ecosystem risk assessment of polycyclic aromatic hydrocarbons in the Luan River, China.* Ecotoxicology, 2010. **19**: p. 827-837.

69. Li, Y., et al., *Air-water exchange and distribution pattern of organochlorine pesticides in the atmosphere and surface water of the open Pacific ocean.* Environmental Pollution, 2020. **265**: p. 114956.

70. Hajir, S., et al., *The association of urinary metabolites of polycyclic aromatic hydrocarbons with obstructive coronary artery disease: A red alert for action.* Environmental Pollution, 2021. **272**: p. 115967.

71. Omar, N.Y.M., et al., *Levels and distributions of organic source tracers in air and roadside dust particles of Kuala Lumpur, Malaysia.* Environmental Geology, 2007. **52**: p. 1485-1500.

72. Sah, R., A. Baroth, and S.A. Hussain, *First account of spatio-temporal analysis, historical trends, source apportionment and ecological risk assessment of banned organochlorine pesticides along the Ganga River.* Environmental Pollution, 2020. **263**: p. 114229.

73. Liu, W.-X., et al., *Aquatic biota as potential biological indicators of the contamination, bioaccumulation and health risks caused by organochlorine pesticides in a large, shallow Chinese lake (Lake Chaohu).* Ecological Indicators, 2016. **60**: p. 335-345.

74. Ozkoc, H.B., G. Bakan, and S. Ariman, *Distribution and bioaccumulation of organochlorine pesticides along the Black Sea coast.* Environmental geochemistry and health, 2007. **29**: p. 59-68.

75. Channa, K., et al., *Prenatal exposure to DDT in malaria endemic region following indoor residual spraying and in non-malaria coastal regions of South Africa.* Science of the total environment, 2012. **429**: p. 183-190.

76. Abdollahdokht, D., et al., *Pesticide exposure and related health problems among farmworkers' children: a case-control study in southeast Iran.* Environmental Science and Pollution Research, 2021. **28**(40): p. 57216-57231.

77. Ferencz, L. and A. Balog, *Pesticides masked with cyclodextrins–a survey of soil samples and computer aided evaluation of the inclusion processes.* Fresen. Environ. Bull, 2010. **19**(2): p. 172-179.

78. Gong, X., et al., *Recording and response of persistent toxic substances (PTSs) in urban lake sediments to anthropogenic activities.* Science of The Total Environment, 2021. **777**: p. 145977.

79. Kucuker, H., et al., *Fatal acute endosulfan toxicity: a case report.* Basic & clinical pharmacology & toxicology, 2009. **104**(1): p. 49-51.

80. Tang, J., et al., *The evolution of pollution profile and health risk assessment for three groups SVOCs pollutants along with Beijiang River, China.* Environmental geochemistry and health, 2017. **39**: p. 1487-1499.

81. Wang, S., et al., *Legacy and emerging persistent organic pollutants in the marginal seas of China: Occurrence and phase partitioning.* Science of The Total Environment, 2022. **827**: p. 154274.

82. Qadeer, A., et al., *Historically linked residues profile of OCPs and PCBs in surface sediments of typical urban river networks, Shanghai: Ecotoxicological state and sources.* Journal of cleaner production, 2019. **231**: p. 1070-1078.

83. Zaghden, H., et al., *Occurrence, origin and potential ecological risk of dissolved polycyclic aromatic hydrocarbons and organochlorines in surface waters of the Gulf of Gabès (Tunisia, Southern Mediterranean Sea).* Marine Pollution Bulletin, 2022. **180**: p. 113737.

84. Wei, Z., et al., *Development History of the Numerical Simulation of Tides in the East Asian Marginal Seas: An Overview.* Journal of Marine Science and Engineering, 2022. **10**(7): p. 984.

85. Zheng, L.-w. and W.-d. Zhai, *Excess nitrogen in the Bohai and Yellow seas, China: Distribution, trends, and source apportionment.* Science of the Total Environment, 2021. **794**: p. 148702.

86. Zheng, W., L. Zou, and Z. Han, *Genetic analysis of the populations of Japanese anchovy Engraulis japonicus from the Yellow Sea and East China Sea based on mitochondrial cytochrome b sequence.* Biochemical Systematics and Ecology, 2015. **58**: p. 169-177.

87. Ya, M., et al., *Impacts of seasonal variation on organochlorine pesticides in the East China Sea and Northern South China Sea.* Environmental Science & Technology, 2019. **53**(22): p. 13088-13097.

88. Lin, T., et al., *Organochlorine pesticides in seawater and the surrounding atmosphere of the marginal seas of China: spatial distribution, sources and air–water exchange.* Science of the Total Environment, 2012. **435**: p. 244-252.

89. Cao, F., et al., *Occurrence, spatial distribution, source, and ecological risk assessment of organochlorine pesticides in Dongting Lake, China.* Environmental Science and Pollution Research, 2021. **28**: p. 30841-30857.

90. Wang, X., et al., *Occurrence, source, and ecological risk assessment of organochlorine pesticides and polychlorinated biphenyls in the water–sediment system of Hangzhou Bay and East China Sea.* Marine Pollution Bulletin, 2022. **179**: p. 113735.

91. Zeng, H., et al., *Risk assessment of an organochlorine pesticide mixture in the surface waters of Qingshitan Reservoir in Southwest China.* RSC advances, 2018. **8**(32): p. 17797-17805.

92. Maqsood, R., et al., *Modeling and predicting students' engagement behaviors using mixture Markov models.* Knowledge and Information Systems, 2022. **64**(5): p. 1349-1384.

93. Yin, J., Y. Zhang, and L. Gao, *Accelerating distributed Expectation–Maximization algorithms with frequent updates.* Journal of Parallel and Distributed Computing, 2018. **111**: p. 65-75.

94. Wei, D., *k-quantiles: L1 distance clustering under a sum constraint.* Pattern Recognition Letters, 2017. **92**: p. 49-55.

95. Salem, S.B., S. Naouali, and Z. Chtourou, *A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach.* Computers & Electrical Engineering, 2018. **68**: p. 463-483.

96. Kuang, Z., et al., *Heavy metal (loid) s in multiple media within a mussel mariculture area of Shangchuan Island, China: Partition, transfer and health risks.* Environmental Research, 2022. **211**: p. 113100.

97. Ramazanova, E., et al., *Spatiotemporal evaluation of water quality and risk assessment of heavy metals in the northern Caspian Sea bounded by Kazakhstan.* Marine Pollution Bulletin, 2022. **181**: p. 113879.

98. He, L., H. Tan, and Z.-C. Huang, *Online handwritten signature verification based on association of curvature and torsion feature with Hausdorff distance.* Multimedia Tools and Applications, 2019. **78**: p. 19253-19278.

99. Zhang, J., et al., *An efficient assembly retrieval method based on Hausdorff distance.* Robotics and Computer-Integrated Manufacturing, 2018. **51**: p. 103-111.

100. Bao, L.-J., et al., *China's water pollution by persistent organic pollutants.* Environmental Pollution, 2012. **163**: p. 100-108.

101. Wang, Q., et al., *Organochlorine pesticide concentrations in pooled serum of people in different age groups from five Chinese cities.* Science of The Total Environment, 2017. **586**: p. 1012-1019.

102. Xu, Y., et al., *Distributions, possible sources and biological risk of DDTs, HCHs and chlordanes in sediments of Beibu Gulf and its tributary rivers, China.* Marine pollution bulletin, 2013. **76**(1-2): p. 52-60.

103. Lammel, G., et al., *Organochlorine pesticides and polychlorinated biphenyls along an east-to-west gradient in subtropical North Atlantic surface water.* Environmental Science and Pollution Research, 2017. **24**: p. 11045-11052.

104. Tian, Y., et al., *Seasonal and inter-annual variations of stable isotopic characteristics of rainfall and cave water in Shennong Cave, Southeast China, and its paleoclimatic implication.* Frontiers in Earth Science, 2021: p. 1187.

105. Galbán-Malagón, C.J., et al., *Persistent organic pollutants in krill from the Bellingshausen, South Scotia, and Weddell Seas.* Science of the Total Environment, 2018. **610**: p. 1487-1495.

106. Li, Z., et al., *Atmospheric deposition and air-sea gas exchange fluxes of DDT and HCH in the Yangtze River Estuary, East China Sea.* Journal of Geophysical Research: Atmospheres, 2017. **122**(14): p. 7664-7677.

107. Huang, D.-Y., et al., *Distribution, regional sources and deposition fluxes of organochlorine pesticides in precipitation in Guangzhou, South China.* Atmospheric Research, 2010. **97**(1-2): p. 115-123.

108. He, P. and D.S. Aga, *Comparison of GC-MS/MS and LC-MS/MS for the analysis of hormones and pesticides in surface waters: advantages and pitfalls.* Analytical Methods, 2019. **11**(11): p. 1436-1448.

109. Luo, X., et al., *Polycyclic aromatic hydrocarbons (PAHs) and organochlorine pesticides in water columns from the Pearl River and the Macao harbor in the Pearl River Delta in South China.* Marine Pollution Bulletin, 2004. **48**(11-12): p. 1102-1115.

110. Wurl, O. and J.P. Obbard, *Distribution of organochlorine compounds in the sea-surface microlayer, water column and sediment of Singapore's coastal environment.* Chemosphere, 2006. **62**(7): p. 1105-1115.

111. Hu, L., et al., *Levels and profiles of persistent organic pollutants in breast milk in China and their potential health risks to breastfed infants: A review.* Science of the Total Environment, 2021. **753**: p. 142028.

112. Chien, M.-Y., C.-M. Yang, and C.-H. Chen, *Organochlorine pesticide residue in Chinese herbal medicine.* Journal of Pesticide Science, 2022. **47**(1): p. 30-34.

113. Gschwend, P., et al., *In situ equilibrium polyethylene passive sampling of soil gas VOC concentrations: Modeling, parameter determinations, and laboratory testing.* Environmental Science & Technology, 2022. **56**(12): p. 7810-7819.

114. Iakovides, M., et al., *Evidence of stockpile contamination for legacy polychlorinated biphenyls and organochlorine pesticides in the urban environment of Cyprus (Eastern Mediterranean): Influence of meteorology on air level variability and gas/particle partitioning based on equilibrium and steady-state models.* Journal of Hazardous Materials, 2022. **439**: p. 129544.

115. Barber, J.L., et al., *Hexachlorobenzene in the global environment: emissions, levels, distribution, trends and processes.* Science of the total environment, 2005. **349**(1-3): p. 1-44.

116. Nyihirani, F., et al., *Level, source, and distribution of organochlorine pesticides (OCPs) in agricultural soils of Tanzania.* Environmental Monitoring and Assessment, 2022. **194**: p. 1-20.

117. Aamir, M., et al., *Occurrence, enantiomeric signature and ecotoxicological risk assessment of HCH isomers and DDT metabolites in the sediments of Kabul River, Pakistan.* Environmental Geochemistry and Health, 2017. **39**: p. 779-790.

118. Dasgupta, S., et al., *Deep seafloor plastics as the source and sink of organic pollutants in the northern South China Sea.* Science of the Total Environment, 2021. **765**: p. 144228.

119. Qiao, W., et al., *Microbial communities associated with sustained anaerobic reductive dechlorination of α-, β-, γ-, and δ-hexachlorocyclohexane isomers to monochlorobenzene and benzene.* Environmental Science & Technology, 2019. **54**(1): p. 255-265.

120. Hao, Y., et al., *Atmospheric concentrations and temporal trends of polychlorinated biphenyls and organochlorine pesticides in the Arctic during 2011–2018.* Chemosphere, 2021. **267**: p. 128859.

121. Sari, M.F., et al., *Atmospheric concentration, source identification, and health risk assessment of persistent organic pollutants (POPs) in two countries: Peru and Turkey.* Environmental Monitoring and Assessment, 2020. **192**: p. 1-16.

122. Girones, L., et al., *Spatial distribution and ecological risk assessment of residual organochlorine pesticides (OCPs) in South American marine environments.* Current environmental health reports, 2020. **7**: p. 147-160.

123. Cowman, A.F., et al., *Malaria: biology and disease.* Cell, 2016. **167**(3): p. 610-624.

124. Daily, J.P., *Monoclonal Antibodies—A Different Approach to Combat Malaria*. 2022, Mass Medical Soc. p. 460-461.

125. Greenwood, B. and T. Mutabingwa, *Malaria in 2002.* Nature, 2002. **415**(6872): p. 670.

126. Boualam, M.A., et al., *Malaria in Europe: a historical perspective.* Frontiers in Medicine, 2021. **8**: p. 691095.

127. Who, *World malaria report 2020: 20 years of global progress and challenges.* World malaria report 2020: 20 years of global progress and challenges, 2020: p. 299.

128. Lalloo, D.G., P. Olukoya, and P. Olliaro, *Malaria in adolescence: burden of disease, consequences, and opportunities for intervention.* The Lancet Infectious Diseases, 2006. **6**(12): p. 780-793.

129. Poku-Awuku, A., et al., *Eliminating malaria in southeast Asia requires more attention on adolescent girls.* The Lancet Child & Adolescent Health, 2021. **5**(12): p. 841-843.

130. Strong, K.L., et al., *Patterns and trends in causes of child and adolescent mortality 2000–2016: setting the scene for child health redesign.* BMJ global health, 2021. **6**(3): p. e004760.

131. Ryan, S.J., C.A. Lippi, and F. Zermoglio, *Shifting transmission risk for malaria in Africa with climate change: a framework for planning and intervention.* Malaria Journal, 2020. **19**: p. 1-14.

132. Wilairatana, P., et al., *Prevalence of malaria and scrub typhus co-infection in febrile patients: A systematic review and meta-analysis.* Parasites & Vectors, 2021. **14**(1): p. 1-13.

133. Asia, T.L.R.H.S., *2030–Countdown to malaria elimination in India and southeast Asia*. 2022, Elsevier. p. 100033.

134. Saba, N., W.K. Balwan, and F. Mushtaq, *Burden of Malaria-A Journey Revisited.* Sch J App Med Sci, 2022. **6**: p. 934-939.

135. Hay, S.I., et al., *Estimating the global clinical burden of Plasmodium falciparum malaria in 2007.* PLoS medicine, 2010. **7**(6): p. e1000290.

136. Patouillard, E., et al., *Global investment targets for malaria control and elimination between 2016 and 2030.* BMJ global health, 2017. **2**(2): p. e000176.

137. Kreppel, K., et al., *Impact of ENSO 2016–17 on regional climate and malaria vector dynamics in Tanzania.* Environmental Research Letters, 2019. **14**(7): p. 075009.

138. Liu, J., et al., *Decline in malaria incidence in a typical county of China: Role of climate variance and anti-malaria intervention measures.* Environmental research, 2018. **167**: p. 276-282.

139. Nosrat, C., et al., *Impact of recent climate extremes on mosquito-borne disease transmission in Kenya.* PLoS neglected tropical diseases, 2021. **15**(3): p. e0009182.

140. Brozak, S.J., J. Mohammed-Awel, and A.B. Gumel, *Mathematics of a single-locus model for assessing the impacts of pyrethroid resistance and temperature on population abundance of malaria mosquitoes.* Infectious Disease Modelling, 2022. **7**(3): p. 277-316.

141. Martineau, P., et al., *Predicting malaria outbreaks from sea surface temperature variability up to 9 months ahead in Limpopo, South Africa, using machine learning.* Frontiers in Public Health, 2022. **10**.

142. Mozaffer, F., G.I. Menon, and F. Ishtiaq, *Exploring the thermal limits of malaria transmission in the western Himalaya.* Ecology and Evolution, 2022. **12**(9): p. e9278.

143. Craig, M.H., R. Snow, and D. le Sueur, *A climate-based distribution model of malaria transmission in sub-Saharan Africa.* Parasitology today, 1999. **15**(3): p. 105-111.

144. Abiodun, G.J., et al., *Modelling the influence of temperature and rainfall on the population dynamics of Anopheles arabiensis.* Malaria journal, 2016. **15**(1): p. 1-15.

145. Yamana, T.K. and E.A. Eltahir, *Projected impacts of climate change on environmental suitability for malaria transmission in West Africa.* Environmental health perspectives, 2013. **121**(10): p. 1179-1186.

146. Brugueras, S., et al., *Environmental drivers, climate change and emergent diseases transmitted by mosquitoes and their vectors in southern Europe: A systematic review.* Environmental research, 2020. **191**: p. 110038.

147. Fischer, L., et al., *Rising temperature and its impact on receptivity to malaria transmission in Europe: a systematic review.* Travel medicine and infectious disease, 2020. **36**: p. 101815.

148. Ivanescu, L., et al., *Climate change is increasing the risk of the reemergence of malaria in Romania.* BioMed Research International, 2016. **2016**.

149. Eikenberry, S.E. and A.B. Gumel, *Mathematical modeling of climate change and malaria transmission dynamics: a historical review.* Journal of mathematical biology, 2018. **77**: p. 857-933.

150. Nabi, S. and S. Qader, *Is Global Warming likely to cause an increased incidence of Malaria?* Libyan Journal of Medicine, 2009. **4**(1): p. 9-16.

151. Rodó, X., et al., *Malaria trends in Ethiopian highlands track the 2000 'slowdown'in global warming.* Nature communications, 2021. **12**(1): p. 1555.

152. Ermert, V., et al., *The impact of regional climate change on malaria risk due to greenhouse forcing and land-use changes in tropical Africa.* Environmental health perspectives, 2012. **120**(1): p. 77-84.

153. Ewnetu, Y. and W. Lemma, *Highland Malaria Transmission Dynamics in Space and Time Before Pre-elimination Era, Northwest Ethiopia.* Journal of Epidemiology and Global Health, 2022. **12**(3): p. 362-371.

154. Mordecai, E.A., et al., *Climate change could shift disease burden from malaria to arboviruses in Africa.* The Lancet Planetary Health, 2020. **4**(9): p. e416-e423.

155. Siraj, A., et al., *Altitudinal changes in malaria incidence in highlands of Ethiopia and Colombia.* Science, 2014. **343**(6175): p. 1154-1158.

156. Asale, A., et al., *The combined impact of LLINs, house screening, and pull-push technology for improved malaria control and livelihoods in rural Ethiopia: study protocol for household randomised controlled trial.* BMC Public Health, 2022. **22**(1): p. 1-14.

157. Okonofua, F., R. Ugiagbe, and L. Ntoimo, *Fostering sustainable malaria prevention and elimination in Africa.* African Journal of Reproductive Health, 2022. **26**(1): p. 9-14.

158. Okumu, F., et al., *What Africa can do to accelerate and sustain progress against malaria.* PLOS Global Public Health, 2022. **2**(6): p. e0000262.

159. Tusting, L.S., et al., *Socioeconomic development as an intervention against malaria: a systematic review and meta-analysis.* The Lancet, 2013. **382**(9896): p. 963-972.

160. Gil-Alana, L.A., *Maximum and minimum temperatures in the United States: Time trends and persistence.* Atmospheric Science Letters, 2018. **19**(4): p. e810.

161. Chaccour, C., *Veterinary endectocides for malaria control and elimination: prospects and challenges.* Philosophical Transactions of the Royal Society B, 2021. **376**(1818): p. 20190810.

162. Fornace, K.M., et al., *Achieving global malaria eradication in changing landscapes.* Malaria Journal, 2021. **20**(1): p. 1-14.

163. Plowe, C.V., *Malaria chemoprevention and drug resistance: a review of the literature and policy implications.* Malaria Journal, 2022. **21**(1): p. 104.

164. Ranjha, R. and A. Sharma, *Forest malaria: the prevailing obstacle for malaria control and elimination in India.* BMJ Global Health, 2021. **6**(5): p. e005391.

165. Karuppusamy, B., et al., *Effect of climate change and deforestation on vector borne diseases in the North-Eastern Indian state of Mizoram bordering Myanmar.* The Journal of Climate Change and Health, 2021. **2**: p. 100015.

166. Ndiath, M.O., et al., *Composition and genetics of malaria vector populations in the Central African Republic.* Malaria Journal, 2016. **15**: p. 1-10.

167. Bhattacharya, S., et al., *Climate change and malaria in India.* Current science, 2006. **90**(3): p. 369-375.

168. Bouma, M., *Epidemic malaria in India and the El Nino southern oscillation.* The Lancet, 1994. **344**(8937): p. 1638-1639.

169. Lindsay, S. and M. Birley, *Climate change and malaria transmission.* Annals of Tropical Medicine & Parasitology, 1996. **90**(5): p. 573-588.

170. Ngarakana-Gwasira, E.T., et al., *Assessing the role of climate change in malaria transmission in Africa.* Malaria research and treatment, 2016. **2016**.

171. Parihar, A., et al., *Plant-based bioactive molecules for targeting of endoribonuclease using steered molecular dynamic simulation approach: a highly conserved therapeutic target against variants of SARS-CoV-2.* Molecular Simulation, 2022: p. 1-13.

172. Tonnang, H.E., R.Y. Kangalawe, and P.Z. Yanda, *Predicting and mapping malaria under climate change scenarios: the potential redistribution of malaria vectors in Africa.* Malaria journal, 2010. **9**(1): p. 1-10.

173. Wang, Z., et al., *The relationship between rising temperatures and malaria incidence in Hainan, China, from 1984 to 2010: a longitudinal cohort study.* The Lancet Planetary Health, 2022. **6**(4): p. e350-e358.

174. Mafwele, B.J. and J.W. Lee, *Relationships between transmission of malaria in Africa and climate factors.* Scientific Reports, 2022. **12**(1): p. 14392.

175. Hay, S.I., et al., *Climate change and the resurgence of malaria in the East African highlands.* Nature, 2002. **415**(6874): p. 905-909.

176. Moise, I.K., et al., *Seasonal and geographic variation of pediatric malaria in Burundi: 2011 to 2012.* International journal of environmental research and public health, 2016. **13**(4): p. 425.

177. Paaijmans, K.P., et al., *Influence of climate on malaria transmission depends on daily temperature variation.* Proceedings of the National Academy of Sciences, 2010. **107**(34): p. 15135-15139.

178. Chaves, L.F., et al., *Indian Ocean dipole and rainfall drive a Moran effect in East Africa malaria transmission.* The Journal of infectious diseases, 2012. **205**(12): p. 1885-1891.

179. Garg, A., et al., *Development, malaria and adaptation to climate change: a case study from India.* Environmental management, 2009. **43**: p. 779-789.

180. Santos-Vega, M., et al., *The neglected role of relative humidity in the interannual variability of urban malaria in Indian cities.* Nature communications, 2022. **13**(1): p. 533.

181. Roy, S.S., *Spatial patterns of malaria case burden and seasonal precipitation in India during 1995–2013.* International Journal of Biometeorology, 2023. **67**(1): p. 157-164.

182. Gómez, C.E., B. Perdiguero, and M. Esteban, *Emerging SARS-CoV-2 variants and impact in global vaccination programs against SARS-CoV-2/COVID-19.* Vaccines, 2021. **9**(3): p. 243.

183. Hu, B., et al., *Characteristics of SARS-CoV-2 and COVID-19.* Nature Reviews Microbiology, 2021. **19**(3): p. 141-154.

184. Das, J.K., S. Chakraborty, and S. Roy, *A scheme for inferring viral-host associations based on codon usage patterns identifies the most affected signaling pathways during COVID-19.* Journal of Biomedical Informatics, 2021. **118**: p. 103801.

185. Harvey, W.T., et al., *SARS-CoV-2 variants, spike mutations and immune escape.* Nature Reviews Microbiology, 2021. **19**(7): p. 409-424.

186. Greaney, A.J., et al., *Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition.* Cell host & microbe, 2021. **29**(1): p. 44-57. e9.

187. McCallum, M., et al., *N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2.* Cell, 2021. **184**(9): p. 2332-2347. e16.

188. Sewell, H.F., et al., *Vaccines, convalescent plasma, and monoclonal antibodies for covid-19*. 2020, British Medical Journal Publishing Group.

189. Tang, X., et al., *Evolutionary analysis and lineage designation of SARS-CoV-2 genomes.* Science Bulletin, 2021. **66**(22): p. 2297-2311.

190. Ripoll, J.G., et al., *Convalescent plasma for infectious diseases: historical framework and use in COVID-19.* Clinical Microbiology Newsletter, 2021. **43**(4): p. 23-32.

191. Zhou, D., et al., *Evidence of escape of SARS-CoV-2 variant B. 1.351 from natural and vaccine-induced sera.* Cell, 2021. **184**(9): p. 2348-2361. e6.

192. Boehm, E., et al., *Novel SARS-CoV-2 variants: the pandemics within the pandemic.* Clinical Microbiology and Infection, 2021. **27**(8): p. 1109-1117.

193. Salvatori, G., et al., *SARS-CoV-2 SPIKE PROTEIN: an optimal immunological target for vaccines.* Journal of translational medicine, 2020. **18**(1): p. 222.

194. Satarker, S. and M. Nampoothiri, *Structural proteins in severe acute respiratory syndrome coronavirus-2.* Archives of medical research, 2020. **51**(6): p. 482-491.

195. Wang, Q., et al., *Structural and functional basis of SARS-CoV-2 entry by using human ACE2.* Cell, 2020. **181**(4): p. 894-904. e9.

196. Zhou, T., et al., *Cryo-EM structures of SARS-CoV-2 spike without and with ACE2 reveal a pH-dependent switch to mediate endosomal positioning of receptor-binding domains.* Cell host & microbe, 2020. **28**(6): p. 867-879. e5.

197. Sztain, T., et al., *A glycan gate controls opening of the SARS-CoV-2 spike protein.* Nature Chemistry, 2021. **13**(10): p. 963-968.

198. Henderson, R., et al., *Controlling the SARS-CoV-2 spike glycoprotein conformation.* Nature structural & molecular biology, 2020. **27**(10): p. 925-933.

199. Khateeb, J., Y. Li, and H. Zhang, *Emerging SARS-CoV-2 variants of concern and potential intervention approaches.* Critical Care, 2021. **25**(1): p. 1-8.

200. Cooper, A. and D. Dryden, *Allostery without conformational change: a plausible model.* European Biophysics Journal, 1984. **11**: p. 103-109.

201. Bozovic, O., et al., *Real-time observation of ligand-induced allosteric transitions in a PDZ domain.* Proceedings of the National Academy of Sciences, 2020. **117**(42): p. 26031-26039.

202. Reid, K.M., X. Yu, and D.M. Leitner, *Change in vibrational entropy with change in protein volume estimated with mode Grüneisen parameters.* The Journal of Chemical Physics, 2021. **154**(5): p. 055102.

203. Gerek, Z.N. and S.B. Ozkan, *Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning.* PLoS computational biology, 2011. **7**(10): p. e1002154.

204. Han, B., et al., *SHIFTX2: significantly improved protein chemical shift prediction.* Journal of biomolecular NMR, 2011. **50**: p. 43-57.

205. Boulton, S., et al., *Implementation of the NMR CHEmical shift covariance analysis (CHESCA): a chemical biologist's approach to allostery.* Protein NMR: Methods and Protocols, 2018: p. 391-405.

206. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome research, 2003. **13**(11): p. 2498-2504.

207. Bader, G.D. and C.W. Hogue, *An automated method for finding molecular complexes in large protein interaction networks.* BMC bioinformatics, 2003. **4**(1): p. 1-27.

208. Selvaratnam, R., et al., *The projection analysis of NMR chemical shifts reveals extended EPAC autoinhibition determinants.* Biophysical journal, 2012. **102**(3): p. 630-639.

209. Xue, L.C., et al., *PRODIGY: a web server for predicting the binding affinity of protein–protein complexes.* Bioinformatics, 2016. **32**(23): p. 3676-3678.

210. Krissinel, E. and K. Henrick, *Inference of macromolecular assemblies from crystalline state.* Journal of molecular biology, 2007. **372**(3): p. 774-797.

211. Wang, J., et al., *Mapping allosteric communications within individual proteins.* Nature communications, 2020. **11**(1): p. 3862.

212. Dokholyan, N.V., *Controlling allosteric networks in proteins.* Chemical reviews, 2016. **116**(11): p. 6463-6487.

213. Verkhivker, G., et al., *Dynamic Profiling of Binding and Allosteric Propensities of the SARS-CoV-2 Spike Protein with Different Classes of Antibodies: Mutational and Perturbation-Based Scanning Reveals the Allosteric Duality of Functionally Adaptable Hotspots.* Journal of chemical theory and computation, 2021. **17**(7): p. 4578-4598.

214. Pettersen, E.F., et al., *UCSF Chimera—a visualization system for exploratory research and analysis.* Journal of computational chemistry, 2004. **25**(13): p. 1605-1612.

215. Bhattacharya, D., et al., *3Drefine: an interactive web server for efficient protein structure refinement.* Nucleic acids research, 2016. **44**(W1): p. W406-W409.

216. Gobeil, S.M.-C., et al., *D614G mutation alters SARS-CoV-2 spike conformation and enhances protease cleavage at the S1/S2 junction.* Cell reports, 2021. **34**(2): p. 108630.

217. Xu, W., et al., *Variations in SARS-CoV-2 spike protein cell epitopes and glycosylation profiles during global transmission course of COVID-19.* Frontiers in Immunology, 2020. **11**: p. 565278.

218. Das, R. and G. Melacini, *A model for agonism and antagonism in an ancient and ubiquitous cAMP-binding domain.* Journal of Biological Chemistry, 2007. **282**(1): p. 581-593.

219. Das, R., et al., *cAMP activation of PKA defines an ancient signaling mechanism.* Proceedings of the National Academy of Sciences, 2007. **104**(1): p. 93-98.

220. Mazhab-Jafari, M.T., et al., *Understanding cAMP-dependent allostery by NMR spectroscopy: comparative analysis of the EPAC1 cAMP-binding domain in its apo and cAMP-bound states.* Journal of the American Chemical Society, 2007. **129**(46): p. 14482-14492.

221. Das, R., et al., *Entropy-driven cAMP-dependent allosteric control of inhibitory interactions in exchange proteins directly activated by cAMP.* Journal of Biological Chemistry, 2008. **283**(28): p. 19691-19703.

222. Das, R., et al., *Dynamically driven ligand selectivity in cyclic nucleotide binding domains.* Journal of Biological Chemistry, 2009. **284**(35): p. 23682-23696.

223. McNicholl, E.T., et al., *Communication between tandem cAMP binding domains in the regulatory subunit of protein kinase A-Iα as revealed by domain-silencing mutations.* Journal of Biological Chemistry, 2010. **285**(20): p. 15523-15537.

224. Ramírez-Aportela, E., J.R. López-Blanco, and P. Chacón, *FRODOCK 2.0: fast protein–protein docking server.* Bioinformatics, 2016. **32**(15): p. 2386-2388.

Appendix **A**

List of program Codes

All the codes and data which I have developed is uploaded at a development platform GitHub. I have attached my account Id https:// https://github.com/Bikash2426/RAS

MATLAB code: Parameters estimation

```
% RAS program for estimating unknown 3-parameters
% three unknown parameters: k, k2 and gamma
% setting these three parameters such that the RAS is always
% remain positive, reach to a stady-state
clc
close all
clear
CAT1=1.4*10^(-2); %s-1;
CAT2= 1.2*10^(-2) ; %s-1;
kMAP=(3*10^(10))/60;  %mmHG M-1 s-1
%reading the PARAMETERS file generated using optimization and model
fit
%teachniques
[NUM,TXT,RAW]=xlsread('parestimate_hypertension1_MAPrange.xls');
%change the file name as per requirement
par=NUM;
steady_state=zeros(20,7);%output files of steady states corresponding
to k, k2 and gamma.
%..............................
exp_no=20;
tspan=0:0.01:10000;
L=length(tspan);
```

```
y0=[1.7*10^(-2),2.06*10^(-4),2.7*10^(-7),2.1*10^(-8),4.1*10^(-
8),2.1*10^(-6),100];% initial conditions of differential equations
%ensuring solution in the positive quadrant
for n=1:exp_no
k=par(n,3);%reading from the estimated values of k's from the input
file
k2=par(n,4);%reading from the estimated values of k's from the input
file
pos=0;%positivity index in the program
gamma=par(n,5);%reading from the estimated values of k's from the
input file
[t,y]=ode45(@(t,y)
simul_diffRAS(t,y,CAT1,CAT2,k,k2,kMAP,gamma),tspan, y0);
for i=1:L
for j=1:7
if (y(i,j)>0)
pos=pos+1;
else
pos=0;
end
end
end
if pos==L*7
disp('system is in +ve quad')
else
disp('continue experiment to ensure the system in +ve quad:')
n
end

%%.....................................................................
```

```
%taking the system to a steady state
if pos==L*7
esp=0.5;%staedy errors
state_ind=0;
for i=1:7
for m=1:10 % it is for testing 10 consequitive numbers of each col
if abs(y(L-11+m,i)-y(L-11+m+1,i))< esp
%if y(k,7)>69 && y(k,7)<101
state_ind=(state_ind+1);
else
state_ind=0;
%end
end
end
end
if state_ind==70
disp('system is in steady state')
%disp('MAP is in the reqired range 70-100\n')
steady_state(n,1)=y(L,1);steady_state(n,2)=y(L,2);
steady_state(n,3)=y(L,3);steady_state(n,4)=y(L,4);
steady_state(n,5)=y(L,5);steady_state(n,6)=y(L,6);
steady_state(n,7)=y(L,7);

else
disp('system in NOT in steady state:')
n
break
end

end
```

```
end
xlswrite('SteadyState_hypertension1.xls',steady_state);
```

MATLAB code: Ordinary differential equation

```
function f = simul_diffRAS(t,y,CAT1,CAT2,k,k2,kMAP,gamma )
%parameters known
KAGT=6.3*10^(-7);   % mol/L/s
hAGT=10*3600;            % s
Renin0=2.06*10^(-13);  % mol/L
hRenin=0.25*3600  ;         % s
sRenin=(log(2)/hRenin)*Renin0;
kRenin=(6.44*10^(4)/3600);  % s-1
cRenin=1.7*10^(-14);        %s-1;
CAItoll= 6.7*10^(-3) ; %s-1;
Kf=(4.91*10^(-5))/(3600);  % s^-1
fa=5.04*10^(2-9);   % mol/L

hANGI=0.62;    % s
hANGII= 18 ;    %s
hAT1R_ANGII=1.5; % s
hAT2R_ANGII=1.5 ; % s

ANGII0=21*10^(-9);  % mol/L

% dAGT/DT

f(1,1)=KAGT-cRenin*y(1)-(log(2)/hAGT)*y(1);   %y(1)=AGT

% dRenin/dt
```

```
f(2,1)=sRenin+Kf*(ANGII0-y(4))*(1-(ANGII0-y(4))/fa)-
(log(2)/hRenin)*y(2)-k2*[y(5)*y(6)];  % y(2)=Renin
% d_ANGI/dt
f(3,1)= cRenin*y(1)+kRenin*(y(2)-Renin0)-[CAItoll+log(2)/hANGI]*y(3);
%y(3)= ANGI

%dANGII/DT
f(4,1)=[CAItoll]*y(3)-[CAT1+CAT2+log(2)/hANGII]*y(4);    % y(4)=ANGII

% d(AT1R-ANGII)/dt

f(5,1)=CAT1*y(4)-(log(2)/hAT1R_ANGII)*y(5)-k*y(6);      % y(5)= AT1R-
ANGII

% d(AT2R-ANGII)

f(6,1)=CAT2*y(4)-(log(2)/hAT2R_ANGII)*y(6);     % y(6)=AT2R-ANGII


% dMAP/Dt
f(7,1)=kMAP*y(4)-gamma*y(7);   %y(7)=MAP



end
```

MATLAB code: Parameter Estimation

```
clc
close all
clear
% Model will be fitted against y(7)(MAP-variable in the ODE) only
%we need to estimate r= k, k2, and gamma
y0=[1.7*10^(-2),2.06*10^(-4),2.7*10^(-7),2.1*10^(-8),4.1*10^(-
8),2.1*10^(-6),100];% initial conditions
tspan = 0:0.01:10000;% discritze time
L = length(tspan);
out=zeros(20,5);
%.............................................
% generating MAP values: normal 20 pts: 70-100; hypertension: 100-
170;
% low-pressure: 40-70;
for i=1:20
%map = linspace(70,100,20);%normal
map=  linspace(101,190,20);%hyper-tension
%map = linspace(40,69,20);%low blood pressure
xx= 0.01*rand(1,51)+map(i);%generating only 50 pts for model fitting
mentioned in "RtoODE"
yvals2=xx;
%.............................................
r   =   optimvar('r',3,'LowerBound',0.0001,'UpperBound',25);%  general
range of all r's
myfcn2 = fcn2optimexpr(@RtoODE,r,tspan,y0);
obj2 = sum(sum((myfcn2 - yvals2).^2));
prob2 = optimproblem('Objective',obj2);
```

```
r0.r = [0.0110 0.5519 0.0039]; %initial guess of r0 that takes the
system at a +ve steady state
[rsol2,sumsq2] = solve(prob2,r0);
disp('Sr.No. map sumsq2 k k2 gamma')
%printing the optimization outcome in a excel file
dd = [i map(i) sumsq2 rsol2.r(1) rsol2.r(2) rsol2.r(3)]
out(i,1)=map(i);out(i,2)=sumsq2;out(i,3)=rsol2.r(1);
out(i,4)=rsol2.r(2);out(i,5)=rsol2.r(3);
end
%.................................................
xlswrite('parestimate_hypertension1_MAPrange.xls',out);    %    store
estimate values
```

MATLAB code: Solution of differential equation

```
n%%% Here we solve the system of differential equations
function f = diffun(~,y,r)
%parameters values
KAGT=6.3*10^(-7);   % mol/L/s
hAGT=10*3600;            % s
Renin0=2.06*10^(-13);  % mol/L
hRenin=900 ;         % s
sRenin=(log(2)/hRenin)*Renin0;
kRenin=(6.44*10^(4)/3600);  % s-1
cRenin=1.7*10^(-14);        %s-1;
CAItoll= 6.7*10^(-3) ; %s-1;
Kf=(4.91*10^(-5))/(3600);  % s^-1
fa=5.04*10^(-7);   % mol/L
hANGI=0.62;    % s
hANGII= 18 ;    %s
```

```
hAT1R_ANGII=1.5; % s
hAT2R_ANGII=1.5 ; % s
ANGII0=21*10^(-9);  % mol/L
kMAP=(3*10^(10))/60;  %mmHG M-1 s-1
CAT1=1.4*10^(-2); %s-1;
CAT2= 1.2*10^(-2) ; %s-1;
% d[AGT]/DT
f(1,1)=KAGT-cRenin*y(1)-(log(2)/hAGT)*y(1);   %y(1)=AGT
% d[Renin]/dt

f(2,1)=sRenin+Kf*(ANGII0-y(4))*(1-(ANGII0-y(4))/fa)-
(log(2)/hRenin)*y(2)-r(2)*[y(5)*y(6)];  % y(2)=Renin
% d[ANGI]/dt
f(3,1)= cRenin*y(1)+kRenin*(y(2)-Renin0)-[CAItoll+log(2)/hANGI]*y(3);
%y(3)= ANGI

%[dANGII]/dT
f(4,1)=[CAItoll]*y(3)-[CAT1+CAT2+log(2)/hANGII]*y(4);    % y(4)=ANGII

% d[AT1R-ANGII]/dt
f(5,1)=CAT1*y(4)-(log(2)/hAT1R_ANGII)*y(5)-r(1)*y(6);        %  y(5)=
AT1R-ANGII
% d[AT2R-ANGII]/dt
f(6,1)=CAT2*y(4)-(log(2)/hAT2R_ANGII)*y(6);     % y(6)=AT2R-ANGII
% dMAP/dt
f(7,1)=kMAP*y(4)-r(3)*y(7);   %y(7)=MAP



end
```

MATLAB code: Transient Analysis

```
% Transient Analysis of RAS system
% We considered all the estimated paramerts k,k2 and gamma as inputs
to
% this program that take the system to a non-zero steady state.
% We calculated numerical solutions for each parameter input
% Then calulated stadard devation among renin and angII solution at
each time point
% All the remaining parameters and intial condition remain same that
set in
% all previous code.
clc
close all
clear
CAT1=1.4*10^(-2); %s-1;
CAT2= 1.2*10^(-2) ; %s-1;
kMAP=(3*10^(10))/60;  %mmHG M-1 s-1
%input excel sheet of parameters
[NUM,TXT,RAW]=xlsread('parestimate_hypertension1_MAPrange.xls');
%change the file name as per requirement
par=NUM;
%.............................
exp_no=length(par);
tspan=0:0.01:10000;
L=length(tspan);
y0=[1.7*10^(-2),2.06*10^(-13),2.7*10^(-7),2.1*10^(-8),4.1*10^(-
8),2.1*10^(-6),100];% Initial conditions
%ensuring solution in the positive quadrant
for n=1:exp_no
```

```
k=par(n,3);%reading from the estimated values of k's from the input
file
k2=par(n,4);%reading from the estimated values of k's from the input
file
gamma=par(n,5);%reading from the estimated values of k's from the
input file
%%%%%%%%
[t,y]=ode45(@(t,y)
simul_diffRAS(t,y,CAT1,CAT2,k,k2,kMAP,gamma),tspan, y0);
sol{n}=y;
%xlswrite([strcat('normal_transol',num2str(n)),'.xlsx'],y);
end
renin=zeros(L,exp_no);angII=zeros(L,exp_no);
for n=1:exp_no
renin(:,n)=sol{n}(:,2);
angII(:,n)=sol{n}(:,4);
end
std_renin=zeros(L,1);std_angII=zeros(L,1);
for i=1:L
std_renin(i)= std(renin(i,:));
std_angII(i)=std(angII(i,:));
end
xlswrite('std_renin_hypertension.xlsx',std_renin);
xlswrite('std_angII_hypertension.xlsx',std_angII);
mean_renin=zeros(L,1);mean_angII=zeros(L,1);
for i=1:L
mean_renin(i)= mean(renin(i,:));
mean_angII(i)=mean(angII(i,:));
end
xlswrite('mean_renin_hypertension.xlsx',mean_renin);
```

```
xlswrite('mean_angII_hypertension.xlsx',mean_angII);
```

MATLAB code: RtoODE Function

```
function solpts = RtoODE(r,tspan,y0)
L = length(tspan);
sol = ode45(@(t,y)diffun(t,y,r),tspan,y0);
solpts = deval(sol,tspan);
solpts = solpts(7,L-50:L); %just consider y(7)with last 50 points
end
```

Appendix B

List of Publications

a. Peer Reviewed International Publications:

i) Das JK, **Thakuri B**, MohanKumar K, Roy S, Sljoka A, Sun GQ, Chakraborty A. Mutation-Induced Long-Range Allosteric Interactions in the Spike Protein Determine the Infectivity of SARS-CoV-2 Emerging Variants. *ACS Omega. 2021 Nov 10;6(46):31312-31327. doi: 10.1021/acsomega.1c05155. PMID: 34805715; PMCID: PMC8592041*.

ii) Wang, C., **Thakuri, B**., Roy, A. K., Mondal, N., Chakraborty, A. (2022). Phase partitioning effects on seasonal compositions and distributions of terrigenous polycyclic aromatic hydrocarbons along the South China Sea and East China Sea. *Science of The Total Environment*, *828*, 154430.

iii) Wang, C., Feng, L., **Thakuri, B.,** Chakraborty, A. (2022). Ecological risk assessment of organochlorine pesticide mixture in South China Sea and East China Sea under the effects of seasonal changes and phase-partitioning. *Marine Pollution Bulletin*, *185*, 114329.

iv) Wang, C., **Thakuri, B.*,** Roy, A. K., Mondal, N., Qi, Y., Chakraborty, A. (2023). Changes in the associations between malaria incidence and climatic factors across malaria endemic countries in Africa and Asia-Pacific region. *Journal of Environmental Management*, *331*, 117264.

b. Article in Preprint Repository and Communicated:

i) **Thakuri B**, Das JK, Roy A.K., Chakraborty A.: Mean-arterial pressure maintenance under feedback controls over the circulating renin-angiotensin systems**.**

Appendix **C**

List of Seminar/Conferences/Workshops attained

I. Attended a one Day special lecture entitled **"Mathematical Modelling in Epidemiology"** organised by Department of Mathematics, Sikkim University.

II. Presented a paper entitled: **"Robust regulation of the renin-angiotensin (RAS) system through the receptors AT1 and AT2 feedback interactions" at** International Conference on Nonlinear Dynamics and Applications organised By SMIT during 9-11 March 2022.

III. Presented a paper entitled **"A Radial Basis Function Method for the solution of Partial differential Equations"** at National Conference on Emerging Trends in Advanced Mathematical Sciences & Its Interdisciplinary Areas organized by Department of mathematical Sciences, Bodoland University, Kokrajhar Assam during 21-22 February ,2020.

IV. Presented a paper entitled **"Mean arterial Pressure (MAP) regulation through the interacting component of Renin-angiotensin System(RAS)"** organised by IMBIC at 16^{th} International Conference MSAST 2022 during December 21-23,2022

V. Workshop on "Techniques for Mathematical Optimization" organized by Department of Mathematics Sikkim University in Collaboration with SQC & OR unit ISI, Kolkata during 29-30 march of 2016.

VI. National Conference on “The Application of Mathematics and Statistics in Industrial, Social and Biological Science “organized by Department of Mathematics, Sikkim Govt. College Tadong, Gangtok, Sikkim during March16-17 of 2018

VII. Three-day NBHM Sponsored Pre-INMO workshop at Rangpo SS school, Sikkim organized by Indian National Mathematics Olympiad (Sikkim Chapter) in collaboration with Department of Mathematics, SGC Tadong during 10-12 January of 2018.