DEVELOPMENT OF INVERSE OPTIMIZATION MODEL FOR IDENTIFICATION OF VIRUS SOURCES IN THE SATURATED-UNSATURATED ZONE

Thesis submitted in partial fulfillment of the requirements

for the award of the degree of

DOCTOR OF PHILOSOPHY

in

Civil Engineering

by

Mamata Das
Roll No. 166104035

Under the guidance of

Prof. Rajib Kumar Bhattacharjya
Prof. Suresh A. Kartha

Department of Civil Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039, Assam, India
August, 2023

DEVELOPMENT OF INVERSE OPTIMIZATION MODEL FOR IDENTIFICATION OF VIRUS SOURCES IN THE SATURATED-UNSATURATED ZONE

Thesis submitted in partial fulfillment of the requirements

for the award of the degree of

DOCTOR OF PHILOSOPHY

in

Civil Engineering

by

Mamata Das
Roll No. 166104035

Under the guidance of

Prof. Rajib Kumar Bhattacharjya
Prof. Suresh A. Kartha

Department of Civil Engineering
Indian Institute of Technology Guwahati
Guwahati - 781039, Assam, India
August 2023

भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
Indian Institute of Technology Guwahati
Department of Civil Engineering, Guwahati, Assam 781039

Dr. Rajib Kumar Bhattacharjya
Professor
Email: rkbc@iitg.ac.in

Dr. Suresh A Kartha
Professor
Email: kartha@iitg.ac.in

CERTIFICATE

This is to certify that the work contained in the thesis entitled " **Development of inverse optimization model for identification of virus sources in the saturated-unsaturated zone**" submitted by **Mamata Das (166104035)** to the Department of Civil Engineering, Indian Institute of Technology Guwahati for the award of the degree of Doctor of Philosophy in Civil Engineering is a record of bonafide research work carried out by her under my supervision and guidance. In my opinion, this thesis work has reached the requisite standard, fulfilling the requirements for the degree of Doctor of Philosophy.

The research work in this thesis has not been submitted in part or whole to any other University or Institute for awarding any degree or diploma.

Date:29.08.2023
Place: Guwahati

Prof. Rajib Kumar Bhattacharjya

29/08/2023
Prof. Suresh A. Kartha

DECLARATION

I do hereby declare that the thesis entitled **" Development of inverse optimization model for the identification of virus sources in the saturated-unsaturated zone "** has been conducted by me as part of my Doctor of Philosophy degree in Civil Engineering under the guidance of Prof. Rajib Kumar Bhattacharjya and Prof. Suresh A. Kartha, Department of Civil Engineering, Indian Institute of Technology Guwahati, Assam, India.

I confirm that this research has not been previously submitted to obtain any other degree or diploma at either this institute or any other academic institution. Proper acknowledgement has been provided in cases where the findings are derived from existing literature reviews.

Mamata Das.

Date: 29.08.2023 (**Mamata Das**)

Place: Guwahati

Dedicated to

My Family

ACKNOWLEDGEMENT

I am sincerely grateful to all those who have been instrumental in completing this journey, culminating in successfully submitting this Ph.D. thesis. Your support, guidance, and encouragement have been invaluable and deeply appreciative.

First and foremost, I am profoundly thankful to my supervisors, Prof. Rajib Kumar Bhattacharjya and Prof. Suresh A Kartha, for their invaluable mentorship, expertise, and constant encouragement. Their insightful feedback, tireless dedication, and willingness to share their knowledge have been crucial in shaping the direction and quality of this research.

I wish to acknowledge the chairman, Prof. Arup Kumar Sarma, and the members, Prof. Sreeja P and Prof. Karuna Kalita, of my doctoral committee for their rigorous examination of this thesis. Their constructive assessment and valuable suggestions have enhanced the rigour and depth of my research.

I am indebted to my seniors, Dr. Leichombam Sophia, Dr. Jagadish Talukdar, Dr. Rajeev Gandhi, and Dr. Brigumani Sharma, whose experiences and insights gave me a solid foundation to build my research. Their willingness to share their knowledge and lessons learned has shaped my work.

I would also like to thank Dr. Aparimita Priyadarshini, Dipsikha Devi, Amrutha Suresh, Prajna Parmita, Arghya Ghosh, Juhi Dhuriya, Sayan Haldar and all my lab mates for their encouragement and support.

Lastly, I thank my father, Mr. Nitai Chandra Das; my mother, Ms. Putuli Das; my Husband, Mr. Bikash Bhuyan; my sister, Dr. Papiya Das; my brother-in-law, Dr. Parikshit Kakati and my younger sister, Neelam Das for accommodating me and supporting my research endeavour.

(Mamata Das)

ABSTRACT

Groundwater aquifers, which are essential drinking water sources, are vulnerable to virus contamination. Virus contamination in groundwaters can occur through various pathways, including septic systems, landfills, agricultural activities, and sewage discharge. Once introduced into the groundwater, viruses can survive for extended periods, potentially leading to widespread contamination of drinking water sources. Viruses can cause various illnesses, including gastrointestinal infections, respiratory infections, and neurological disorders in humans. As such, virus contamination in groundwater can pose a significant risk to human health. Identifying the source of viruses in groundwater can help determine the appropriate measures needed to prevent or mitigate contamination, protect public health, and maintain the quality of groundwater resources. This study focuses on developing an effective source identification model in an unconfined groundwater aquifer considering both the unsaturated and saturated zones. As such, a source identification model requires information about the spatial and temporal distribution of the virus in the aquifer. The existing commercial computational models, such as HYDRUS-3D and MT3DMS, are widely used for simulating contaminant transport in groundwater. For example, MT3DMS cannot account for variable degrees of saturation, and HYDRUS 3D cannot handle conditions such as variations in activation and inactivation rates of viruses due to changes in subsurface moisture content, temperatures, etc. This research develops-

- **A novel computational virus transport model to elucidate the migration of virus contamination from the unsaturated zone to the saturated area, addressing limitations in commercial models by accounting for input parameter variations dependent on saturation degree.** The model has been validated using experimental data to assess its ability to generalize and accurately predict virus transport in scenarios.
- In the virus transport phenomenon, several key parameters are commonly considered that describe the behavior and characteristics of the virus in the environment. However, some of these parameters are not available on field and laboratory scale and thus need a parameter estimation model. The existing models are not designed to perform the parameter estimation for both the zones in a coupled manner. **A parameter estimation model is developed for the coupled unsaturated and saturated zone to estimate the flow and transport parameters using a Metaheuristic algorithm.** The algorithm can estimate flow and transport parameters with minimum relative error.

- Once the computational virus transport model is developed, it is linked with an optimization model to identify the virus sources in the aquifer. The objective function of the optimization model is to minimize the error between the field observed and computationally simulated virus concentration. In the case of virus source identification, the existing models primarily focus on locating sources in the saturated zone of groundwater. However, in practice, virus sources frequently exist in the unsaturated zone, situated above the water table where soil or rock retains partial saturation with water. **The research focused on developing a source identification model that can detect the source location in both saturated and unsaturated zones**. The Shuffled Frog Leaping Algorithm (SFLA) is used as an optimization algorithm to identify the virus source location and source strength.
- The SFLA based source identification model accurately determined the virus source locations and the source fluxes However, the computational time required for the process was quite high. To overcome this limitation **an improved source identification model algorithm is developed by using Recurrent Neural Network (RNN) in conjunction with the computational simulation model and shuffled Frog Leaping Algorithm.** This model is efficient both in computation time and predicting performance.

Keywords: Source location; parameters; optimization; Metaheuristic.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

a	Specific air–water interfacial area.
a_0	Specific interfacial area corresponding to residual saturation of the medium.
C_a	Virus concentration sorbed onto Air-water-Interfaces (AWI).
C_s	Virus concentration sorbed onto Soil-water-Interfaces (SWI).
C_W	The centration of free virus per unit volume in water.
D	The dispersion coefficient.
D^*_{ij}	Molecular diffusion.
D^m_{ij}.	Mechanical dispersion.
g^t	Current best solution in Flower pollination algorithm.
H	Hydraulic head.
h	Pressure head.
h_0	Parameter depending upon the Specific storage (S_S).
h_a	Air entry pressure.
L	Step size drawn from Lévy flight distribution function.
K_{att}^s	Rates of attachment in SWI.
K_{att}^a	Rates of attachment in AWI.
$K_{att}^{s,s}$	Rates of attachment in SWI in the saturated zone.
$K_{att}^{s,u}$	Rates of attachment in SWI in unsaturated zone.
K_{det}^s	Rates of detachment in SWI.
K_{det}^a	Rates of detachment in AWI.
$K_{det}^{s,s}$	Rates of detachment in SWI in the saturated zone.
$K_{det}^{s,u}$	Rates of detachment in SWI in unsaturated zone.
K_D^a	Equilibrium distribution coefficient for sorption onto the AWI.
$K(h)$	Unsaturated hydraulic conductivity.
K_r	Relative hydraulic conductivity.
K_s	Saturated hydraulic conductivity.
K_{xx}	Hydraulic conductivity along x directions.
K_{yy}	Hydraulic conductivity along y directions.
K_{zz}	Hydraulic conductivity along z directions.
$\boldsymbol{M}$	Total number of observation wells.
m	Picard's iterative time step.
N	Total number of time steps.

n	Fitting parameter in the moisture retention
n_m	Number of memeplexes in SFLA.
n_f	Virtual frogs within each memeplex in SFLA.
OC_i^j	Observed concentration at i^{th} time step for j^{th} well location;
OH_i^j	Simulated concentration at i^{th} time step for j^{th} well location;
p_i	Probability for the triangular distribution in SFLA.
Q_I	Inflow rate into the cell.
Q_o	Outflow rate from the cell.
Q_e	Pumping or recharge rate.
q	Velocity flux indicates the rate of fluid flow obtained using the flow model.
q_e	Source or sink term.
R	Retardation coefficient.
R^2	Co-efficient of regression
SC_i^j	Simulated concentration at i^{th} time step for j^{th} well location;
Sf	The source strength of virus contamination represents the amount of virus released into the system.
SH_i^j	Simulated hydraulic head at i^{th} time step for j^{th} well location obtained from the numerical simulation model.
S_i	Sobol's First order sensitivity indices.
S_{ij}	Sobol's Second-order indices
Sl	Virus source location
Ss	Specific storage
S_t	Sobol's Total-order indices.
S_w	Degree of saturation
t	Time steps
v	Velocity vector
θ	Water content
$\boldsymbol{\theta_a}$	Air content.
θ_r	Residual water content.
θ_s	Saturated water content.
ΔV	Volume of the cell.
$V(y)$	Total variance decomposition in sobol's method.
ρ_b	Bulk density of soil.

μ_w	Inactivation rate coefficients associated with water.
μ_s	Inactivation rate coefficients associated with SWI.
μ_a	Inactivation rate coefficients associated with AWI.
$\boldsymbol{\gamma_s}$	Rates of adsorption to SWI.
$\boldsymbol{\gamma_a}$	Rates of adsorption to AWI.
α	Soil parameter.
α_{ijk}	Fourth-order dispersivity tensor.
α_L	longitudinal dispersivity.
α_{TH}	Transverse Horizontal Dispersivity.
α_{TV}	Transverse vertical dispersivity.
ω	Spatial weighting factors are used when evaluating a dispersion cross-term.
$\Gamma(\lambda)$	Standard gamma function.

LIST OF ABBREVIATIONS

AARE	Average Absolute Error
ANN	Artificial Neural Network
AWI	Air- water Interfaces
FDM	Finite Difference Method.
FEM	Finite Element Method
FEFLOW	Finite Element subsurface FLOW system
FOI	First Order Indices
FPA	Flower Pollination Algorithm.
FVM	Finite Volume Method (),).
GA	Genetic Algorithm
GWT	Ground Water Table
MODFLOW	Modular three-dimensional finite-difference ground-water flow model
MT3DMS	Modular three-dimensional multispecies transport model
MSE	Mean Square Error
NSM	Numerical Transport Simulation Model
OC	Observed Concentration
OH	Observed Head
PCGM	Preconditioned Conjugate Gradient Method
RE	Relative Error
RHS	Right Hand Side
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
SC	Simulated Concentration
SFLA	Shuffled Frog Leaping Algorithm
SH	Simulated Head
SOI	Second Order Indices
SWI	Soil- water Interfaces

CHAPTER 1
INTRODUCTION

1.1 General:

Global population growth is increasing the demand for freshwater at an alarming rate. About 97% of Earth's water is in oceans, and only 3% are the freshwaters. Of the small amount of freshwater on Earth, 69% is in glaciers, 30% is in groundwater, and less than 1% is in surface water. Due to the lack of fresh surface water in some areas, people rely primarily on groundwater. However, rapid industrialization and urbanization of the modern world have led to increased groundwater contamination. Among different groundwater contaminants, microbial contamination, such as viruses, can cause diseases and pose serious health risks to humans (Craun, 1992). The primary sources of viruses in an aquifer are leakage from septic tanks, sanitary landfills, and cesspools (Yates, 1985). Many of these sources are found in unsaturated zones, which makes it crucial to understand how these microbes are transported into the subsurface domain. Usually, it takes longer for groundwater contamination in an aquifer to move from one location to another. However, due to rapid pumping, this contaminant moves toward the well location and poses a severe threat to people who use it for drinking water (Almasri and Kaluarachchi, 2004, Taghavi et al., 1994, Willis and Yeh, 1987). So, to take appropriate remedial measures, it is essential to identify the source of virus pollution as early as possible.

1.2 Viruses in Groundwater:

Viruses are small, contagious agents that replicate only in the living cells of other organisms. Viruses are composed of nucleic acids, either DNA or RNA, enclosed within a protective protein coat known as a capsid, where each capsomere is made of an identical protein subunit. The proteins may also be surrounded by a lipid membrane called an envelope. It does not reproduce through cell division since it is a cellular virus. Instead, they replicate within a host cell, using its metabolic process to make multiple copies of themselves. Viruses can range in size from 20 nanometres to 300 nanometres, which is much smaller than bacteria. While some filoviruses measure up to 1400nm in length, their diameters are only about 80nm. In addition to animals and plants, the virus infects microorganisms, such as bacteria and archaea (Koonin et al., 2006). Viruses can spread in many ways: insects transmit viruses to plants, coughs and sneezes spread viruses to people, food and water transmit viruses, and body fluids transfer

viruses. Many waterborne infectious diseases are transmitted through groundwater around the world. Human enteric viruses are the most common agents of contamination of groundwater. Hepatitis, Polio, Coxsackie, Adenovirus, Rotavirus, and Norwalk-like viruses are some of the most common pathogenic viruses that can be transported on water paths (Keswick & Gerba, 1980). When temperatures are low, and soils are moist, viruses can persist for months in soils and groundwater (Straub et al., 1993). According to many researcher, clay soils enhances virus survival rates more than sandy soils (Parsai et al., 2018).A linear correlation links temperature and viral decay, indicating a clear association between virus survival and temperature (Gerba & Choi., 2006; Pinon & Vialette., 2019).

1.2.1 Sources of virus in groundwater:

Viruses enter the aquifer through many sources, including leakage from septic tanks, sewer lines, and sanitary landfills (Yates, 1985). Water supply and storm drainage leaks contribute significant recharge, but sewage system leaks, septic tank issues, untreated wastewater injection, and accidental spills can introduce harmful pathogens or viruses to urban aquifers (Gotkowitz et al., 2016; Wakode et al., 2018). The viruses enter through the unsaturated zone or the fractured rocks and move to the fresh groundwater aquifer. Viruses in groundwater do not get multiplied inside the groundwater alone. They replicate by themselves when they find a host in the form of a living cell.

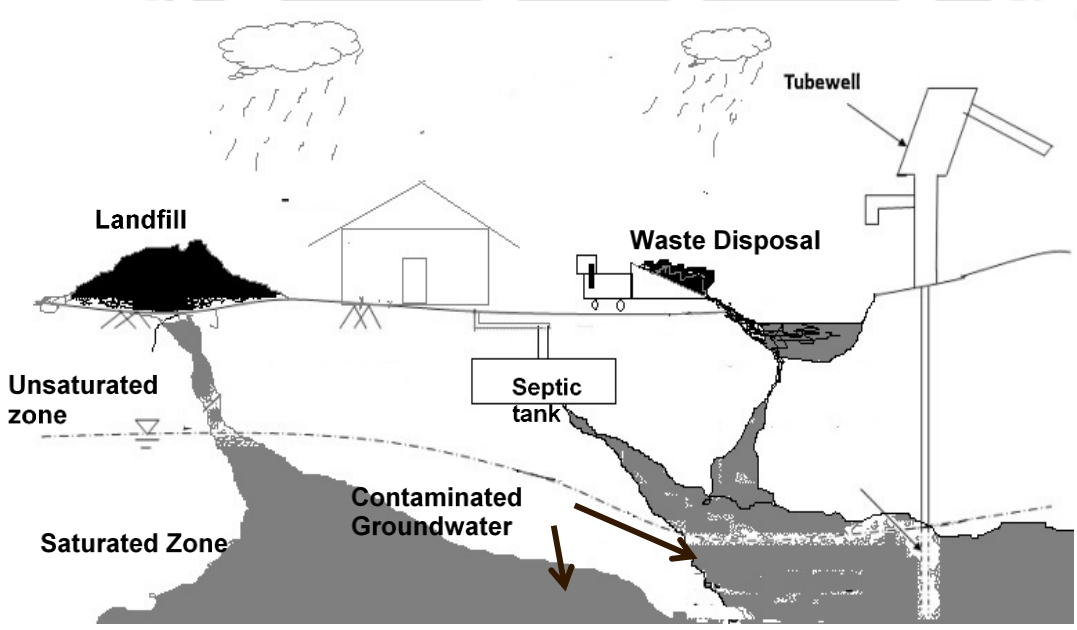


Figure 1.1:Sources of virus contamination in groundwater.

There are many factors on which the survival of the virus in groundwater depends, which will be described later in this chapter. Among them, the size of the virus is an essential factor because due to its tiny size (20-350 nm), it can travel very long distances if the conditions are favorable and most disinfection resistant nature (Loveland et al., 1996). After several investigations, it is confirmed that the virus travels from a contamination source through the unsaturated soil to reach the groundwater (Lance & Gerba, 1984). Figure 1.1 shows the sources of different groundwater contamination and its movement to groundwater.

1.2.2 Survival of virus in groundwater aquifer:

As viruses in groundwater do not get multiplied inside the groundwater alone, they replicate by themselves only when they find a host in the form of a living cell. Otherwise, they will become inactive due to the groundwater's pH level or the environment's temperature. As viruses are smaller, they can travel very long paths under favorable conditions and in most disinfection-resistant nature (Loveland et al., 1996). Researchers often employ spiked viruses to contaminate water artificially for studying viral survival. However, this method may not accurately represent the actual threats in groundwater environments (Pinon & Vialette., 2019). The pH, temperature, organics, ionic strength, and cations, among others, may affect the virus' survival in groundwater (Gassilloud & Gantzer., 2005; Hunt & Johnson., 2017).It is challenging to predict virus survival because these factors vary tremendously over time. The persistence of viruses in groundwater is estimated to be at least 28 days (Wellings et al., 1975). Since the effect of sunlight is eliminated in groundwater and the temperature is lowered, a virus can survive for a long time. Biological studies predict that evaporation and temperatures are high near the ground surface. This causes aerobic soil microorganisms to contribute to the rapid inactivation of viruses, unlike the virus present in the deep groundwater aquifer. In this way, virus survival in the soil is affected by various factors, including survival rates and their continuation.

1.3 Movement of virus in groundwater aquifer:

It is necessary to understand virus movement from the sources to the groundwater aquifer to develop any groundwater management model or to propose any remedial measure to control groundwater contamination due to viruses. The transport process of viruses in an aquifer can be developed by considering advection, dispersion, adsorption, and inactivation mechanisms.

1.3.1 Advection-dispersion:

Advection causes translation of the solute field by moving the solute with the bulk fluid flow velocity. Dispersion causes the 'spreading' of the solutes into the groundwater. It is composed of both molecular and mechanical dispersion. As mentioned by Yates and Yates (1988), the advection-dispersion expression is given in index notations by equation (1.1).

$$\frac{\partial C_w}{\partial t} = \frac{\partial}{\partial x_i}\left[D_{ij} \frac{\partial C_w}{\partial x_j} - v_i C_w \right] \tag{1.1}$$

where C_w = concentration of viruses in aqueous phase, v_i = pore water velocity in the corresponding flow direction, D_{ij} = hydrodynamic dispersion coefficient, t = time period, and i and j are suffixes of index notations. As evident in equation (1.1), i is the dummy index, and j is the free index. The transport of viruses in groundwater is different from other solute tracers because it is governed by two other processes, which are adsorption and inactivation.

1.3.2 Adsorption:

Adsorption is a physical or chemical process by which the solute attaches to the soil and air interfaces. As a result of the adsorption, the movement of the solute or mixing of solute becomes much slower than the movement of groundwater in an aquifer. Thus, sorption induces a retardation effect on the solute movement in groundwater aquifers. During the virus transport, irreversible and reversible adsorption occurs. In the case of irreversible sorption, attachment occurs with no detachment, termed equilibrium sorption. Whereas in reversible sorption, both attachment and detachment occur, termed kinetic sorption. In the saturated zone, virus sorption is significantly affected by aqueous and liquid-solid interfaces. The unsaturated zone comprises three phases, and the air-liquid interfaces play a significant role in the virus adsorption process in the aquifer (Wan & Wilson, 1992). The equation (1.2) gives the transport equation accounting for the adsorption process.

$$\frac{\partial C_w}{\partial t} = \frac{\partial}{\partial x_i}\left[D_{ij} \frac{\partial C_w}{\partial x_j} - v_i C_w \right] - \frac{\rho_b}{\eta} \frac{\partial S}{\partial t} \tag{1.2}$$

where ρ_b is the bulk density of the soil matrix, S is the sorbed concentration of the virus, and η is the porosity of the soil matrix.

1.3.3 Inactivation:

Inactivation is a chemical process by which the viruses become inactive or unable to infect. Virus inactivation is dependent on many factors, such as pH, temperature, organics, ionic strength, cations, etc. The inactivation rate is denoted by μ. The transport equation considering the inactivation process is given by equation (1.3).

$$\frac{\partial C_w}{\partial t} = \frac{\partial}{\partial x_i}\left[D_{ij}\frac{\partial C_w}{\partial x_j} - v_i C_w \right] - \frac{\rho_b}{\eta}\frac{\partial S}{\partial t} - \mu C_w \qquad (1.3)$$

The inactivation and adsorption of viruses onto the soil matrix play a significant role in distinguishing virus transport from other subsurface solute transport (Vilker & Burge, 1980). Virus inactivation rates differ considerably for liquid and sorbed phases (Hurst et al., 1980;Jin & Flury, 2002). It has been demonstrated in several studies that virus inactivation depends on time (Chrysikopoulos & Sim, 1996); Chrysikopoulos & Vogler, 2004; Ojha et al., 2011). Because of their resistance to protein coat disruption and nucleic acid degradation, sorbed viruses on soil can survive for a long time (Yates 1980). Viral sorption is influenced by ionic strength, pH, and virus properties (Joshi et al., 2013 ; Jin & Flury, 2002). There are two types of virus sorption: equilibrium sorption and non-equilibrium sorption. Non-equilibrium sorption results when virus attachment and detachment occur slower than bulk fluid velocity, whereas equilibrium sorption occurs when attachment and detachment occur faster (Joshi et al., 2013). In 1991, Bales et al. reported that equilibrium sorption parameters may be misrepresented in non-equilibrium conditions. In unsaturated systems, due to the air-water interface, additional sorption takes place compared to saturated systems (Jin et al., 2000). Therefore, to understand the virus transport process from the unsaturated zone to the groundwater aquifer, it is necessary to develop a non-equilibrium model considering inactivation and sorption in liquid as well as air-liquid and liquid-solid interfaces. To develop the virus transport model in the unsaturated-saturated zone, one must develop the flow model for the proposed study domain before modeling the virus movement from the unsaturated zone to the groundwater aquifer. This is because, in an unsaturated zone, the moisture velocity and degree of saturation vary with space and time.

1.4 Parameter estimation model:

One must consider the necessary flow and transport parameters to develop an accurate numerical simulation model with proper boundary conditions. Nevertheless, there are a few

flow and transport parameters that cannot be measured on the field or laboratory scale. In such cases, the parameter estimation technique plays a significant role in identifying those parameters. The linked-simulation optimization model has proven effective in identifying both flow and transport parameters in porous media (Eching & Hopmans, 1993; Jacques et al., 2002; Kodešová et al., 1998; Kool et al., 1987;Inoue et al., 2000; Šimůnek & Van Genuchten, 1996; Wildenschild et al., 2001). An inverse optimization technique aims to minimize the difference between the simulated and observed values. Numerous researchers have used it to solve various problems related to groundwater management (Borah & Bhattacharjya, 2013; Borah & Bhattacharjya, 2015). However, only a few studies estimated virus transport parameters in the groundwater aquifer.

1.5 Importance of source identification model:

Groundwater contamination may arise from a point or non-point sources. A point source originates from a small scale, whereas a non-point source refers to large-scale sources that get diffused from many smaller sources. Contaminant sources are always an issue when modeling or managing contaminant transport at the field scale. Modelers are typically presented with information about a particular contaminant in some observation wells. The modeler must identify the source and location of the contaminant based on the given data. A problem like this is known as a source identification problem (Leichombam & Bhattacharjya, 2016). The unknown groundwater pollution sources can be identified using different contaminant transport inverse models such as Optimization, probabilistic, analytical, regression, and direct approaches. The problem is solved by minimizing the differences between the observed and simulated concentrations using an arbitrary value as a contaminant source location and its strength. It is, however, essential to address issues such as finding the source with limited observation data and early detecting viruses in the aquifer during the source identification process. Therefore, an active source can be inactivated and stop their transport to the groundwater aquifer by accurately identifying the contaminant sources.

1.6 Motivation for the Study:

In many places in the world, microbial contamination in drinking water has created serious health hazards to human beings. Over the last ten years, the United States Government has focused on controlling groundwater contamination. They have adopted different removal strategies for viruses from the groundwater. Although it is difficult, many projects are underway in the Netherlands and northern America to control microbial contamination.

Microbial contamination of drinking water causes cholera, diarrhea, gastroenteritis, and other diseases. More than half of the 3500 inhabitants of a Karst village in Switzerland suffered gastroenteritis in 1998 due to Campylobacter and Listeria organisms in the area. This was caused because a failed pump blocked the sewerage, infiltrating sewage into the groundwater. Thus, Fecal Coliforms have been found in groundwater collected for drinking purposes (Häfliger et al., 2000). In 2000, Walkerton, Ontario, experienced one of the most tragic groundwater contamination stories. During this outbreak, 7 people died, and more than 2,300 were sickened by pathogenic E coli (D. O'Connor, 2002; D. R. O'Connor, 2002).There were approximately 5800 cases of diarrhea, abdominal cramps, vomiting, and headaches in North Battleford, Saskatchewan, in 2001 due to Cryptosporidium parvum. After extreme precipitation events in May–July 2004, coliform and Escherichia coli were transported from wastewater treatment septic tanks and thus contaminated groundwater on South Bass Island, Ohio. The city reported approximately 1,450 cases of gastroenteritis and campylobacteriosis. These are a few outbreaks that have been caused due to pathogenic contamination in groundwater aquifers.

Like India, there are many developing countries where the proper sewer system is not yet developed. Thus, most people use an isolated septic tank for sanitation. The leakages from septic tanks and sewer lines are one of the significant sources of virus contamination in groundwater aquifers. Such groundwater contamination directly affects the lives of people with serious health issues like diarrhea, hepatitis, gastrointestinal diseases, etc. In 2011, a study was carried out by the Central Groundwater Board, Faridabad, to check the bacteriological content in the drinking water. They collected a few groundwater samples from the Amravati district, Maharashtra, and it was reported that the presence of faecal coliform bacteria in groundwater. There may be many places in India where groundwater is affected by different microbial organisms. Therefore, taking proper remedial measures to control groundwater contamination by pathogenic viruses becomes imperative. Since remediation of the contaminated aquifer is complex and the contaminants cannot be removed entirely, it is adequate to identify the sources of such contamination and make it inactive. As mentioned previously, the sources of viruses are mainly present in the unsaturated zone, so this study aims to identify the sources of such virus contamination in groundwater aquifers.

1.7 Research Objective:

This research aims to develop an effective source identification model to identify the virus sources in the aquifer. As mentioned in the above sections, the sources of virus contamination

are mainly present in unsaturated zones, so this study developed a source identification model to detect the sources present in unsaturated and saturated zones. The model is proposed to identify the virus source location in a three-dimensional orthogonal cartesian coordinate space. In order to fulfill the research mentioned above, this thesis work has formulated the below-mentioned research objectives. The outline of the whole thesis work is given in Figure 1. 2.

1. Three-dimensional virus transport model in an unconfined aquifer considering both the unsaturated and saturated zones.
2. An effective parameter estimation model to estimate the flow and virus transport parameters. Global sensitivity analysis of both the flow and transport parameters.
3. Inverse optimization model to identify the unknown virus sources in unconfined aquifer considering source locations, source fluxes, and source strength as variables.
4. Development of an improved unknown virus source identification model using Numerical and RNN-based simulation-optimization tools.

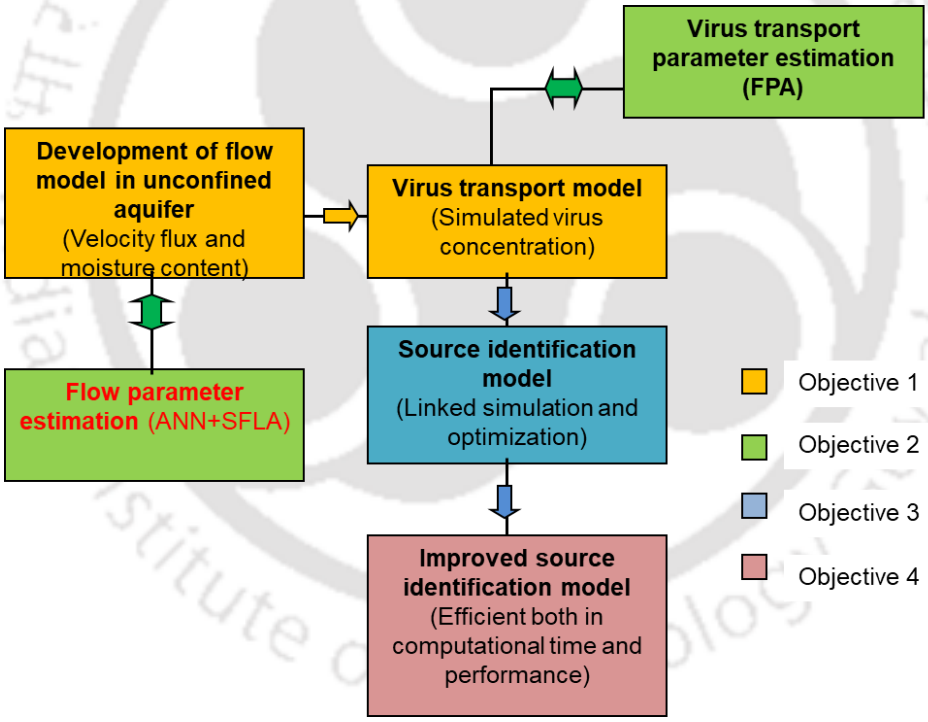


Figure 1. 2:Flowchart showing the overview of the research work.

1.8 Thesis orientation:

This thesis includes seven chapters, from the introduction to the conclusion chapters, and a summary of the work. A brief overview of each of the chapters is given as follows.

Chapter 1: This chapter briefly discusses the importance of groundwater and the reason for groundwater contamination. The chapter concisely mentioned groundwater contamination due

to viruses as a contaminant, its survival, and its movement in groundwater aquifers. The motivation for the research work is also presented in this chapter and includes the research objectives.

Chapter 2: This chapter gives a detailed discussion of the previous research works on flow and virus transport modeling in both saturated and unsaturated zones. It has also reported the details of the advancement of parameter estimation and source identification techniques over recent years. Finally, this chapter concludes with the research gap.

Chapter 3: In this chapter, the three-dimensional virus transport model is developed in an unconfined aquifer, considering both the saturated and unsaturated zones. The chapter consists of two parts- In the first part, the three-dimensional flow model is developed, and the solution obtained from the flow model is incorporated in the second part, where the virus transport model is developed. The model has been validated with different experimental and numerical models available in the literature.

Chapter 4: This chapter comprises the effective parameter estimation model for estimating the hydraulic and transport parameters. These are mainly two inverse optimization models for estimating the flow and virus transport model. This chapter ends with a global sensitivity analysis to understand the relative importance of each flow and transport parameters with the model output.

Chapter 5: In this chapter, the virus source identification model is developed with the information from the observed data of the virus concentration in the aquifer. The model is developed for two scenarios, one considering the source location to be present in the unsaturated zone, and in the second scenario, one source is considered in the saturated zone and another in the unsaturated zone. The model assumed unknown source location, source strengths, and inflow source flux.

Chapter 6: This chapter is mainly the improvement in the source identification model as developed in the previous chapter. This chapter proposed a new methodology using an approximate simulator-numerical model optimization algorithm for achieving efficiency in computational time and predicting performance in identifying the virus sources.

Chapter 7: In this chapter, the whole research work is summarized and briefly presented with all the conclusions obtained from the objectives. Lastly, this chapter ends with the future scope of the study and the possibility of application of the proposed strategy in the field.

CHAPTER 2
REVIEW OF LITERATURE

2.1 Overview:

In the last few decades, several studies have been carried out to understand groundwater contamination ever since there were outbreaks of epidemics due to virus contamination. In this chapter, the work carried out by different researchers to study the transport of viruses in groundwater, the parameters used, and its estimation and sensitivity to the model output, its survival, and pollution source identification are narrated briefly.

2.2 Virus as a contaminant in Groundwater:

Groundwater contamination due to pathogenic microorganisms has become a severe problem. By surface runoff, the pathogenic organisms get transported to streams, penetrating through the unsaturated zone and finally reaching the groundwater (Gerba, 1984). (Mack et al., 1972; Wellings et al., 1971) demonstrated that viruses could be transmitted for long distances via soil to groundwater. Most groundwater viruses are animal viruses that may infect humans, such as adenovirus, coliphage, coxsackievirus, enterovirus, hepatitis, poliovirus, and rotavirus (Keswick & Gerba, 1980; Yates & Yates, 1988). Several laboratory studies were carried out to understand the survival of such viruses in groundwater. According to Wellings et al., (1975) viruses can survive in groundwater for at least 28 days, whereas other experiments have shown that viruses can survive for more than 200 days. Thus, assessing the virus's survival in groundwater is difficult as different factors, such as pH value, temperature, ionic strength, etc, influence it. The inactivation of the virus in groundwater depends on subsurface temperature (Yates & Ouyang, 1992). It remains infective much longer at a lower temperature (1-8 ^{0}C) than at a higher temperature (20-32 ^{0}C) (Park et al., 1992). Therefore, it can be assumed that viruses that penetrate the soil surface are expected to survive for prolonged periods compared to those retained near the soil surface. Thus, if these contaminated water supplies are adequately treated before it is used, public health may be safe. In order to control the virus contamination in groundwater, this research aims to detect the source location of the virus contamination in the aquifer; thus, identifying the sources and making it inactive will act as an alternative remedial measure to control viral contamination in groundwater. To develop this model, a synthetic groundwater problem is considered with natural boundary conditions to study the movement of virus (as a contaminant) from its sources to the groundwater aquifer. Different virus transport

model in groundwater aquifer considering unsaturated, saturated, and variably saturated zones is available in the literature. Past developed models are mentioned in section 2.3 in this chapter.

2.3 Virus transport model:

In an aquifer, contaminants flow, and transport can be simulated using the governing flow and transport equation. Due to the lack of closed-form solutions for field scale problems, numerical methods are generally used to solve the equations, including Finite Difference Method (FDM), Finite Volume Method (FVM), and Finite Element Method (FEM). As such, several mathematical models are available in the literature. Since the virus sources are also present in unsaturated zones, many researchers have developed mathematical models for unsaturated porous media separately. The unsaturated zone consists of liquid, solid, and air phases. Due to multiple phases, sorption process of virus is significantly affected by liquid-solid and air-liquid interfaces (Wan & Wilson, 1992). In this zone, the virus sorption and inactivation are influenced by the soil moisture content and subsurface temperature fluctuations (Thompson & Yates, 1999; Vilker & Burge, 1980; Vilker, 1981). Therefore, the fluid flow equation is solved to determine the flow velocity and hydraulic heads before solving the transport equation.

The first one-dimensional model VIROTRANS (Tim & Mostaghimi, 1991), developed to study virus transport in unsaturated zones, considers virus sorption process to be in equilibrium and virus inactivation rates are the same for both adsorbed and liquid phase viruses. An alternative model VIRALT was developed to account for the first-order inactivation rate in both steady and transient states (Park et al., 1992). Since temperature is an important factor affecting the inactivation rate, a one-dimensional numerical model VIRTUS was developed to predict the vertical movement of the virus with a temperature-dependent inactivation rate (Yates & Ouyang, 1992). These models involve equilibrium adsorption, but kinetic adsorption mainly governs virus attenuation (Bales et al., 1991, 1997). (Schijven et al., 1999) described virus transport using one-dimensional one-site kinetic sorption models on a field scale. According to them, first-order kinetic rate processes could describe both attachment and detachment of viruses from solid phases. Since air-liquid interfaces had not been considered for adsorption before then, a one-dimensional mathematical model of viruses adsorbing at liquid-solid and air-solid interfaces, along with their inactivation rates, was developed by (Anders & Chrysikopoulos, 2009). By integrating kinetic sorption phenomena, Chu et al., (2001)modified this mathematical model. Their study considered kinetic adsorption to soil-water interfaces and kinetic adsorption to air–water interfaces. There is a significant difference

between the transport of dissolved contaminants and viruses in unsaturated zones since dissolved contaminants do not react with the air-water interface, whereas viruses themselves are colloid particles that may react with the soil-water interface and air-water interface. In order to study the co-transport of colloids and viruses in unsaturated zones under steady-state conditions, a one-dimensional mathematical model was developed(Seetha et al., 2014). Rajsekha et al. (2016) investigate the impact of moisture content, inactivation rate constant, pore velocity, and mass transfer coefficients on virus movement through vertical depth in unsaturated zones. A two-dimensional axisymmetric virus transport model was developed to evaluate the virus removal in an ASR scheme considering aquifer heterogeneity, revealing significant differences compared to homogeneous aquifer models (Torkzaban et. al., 2019). Sasidharan et. al., (2021) used numerical experiments with HYDRUS software and laboratory column data to analyze the affects of subsurface heterogeneity on virus transport when field studies aren't possible. As such steady-state flow simulations were conducted to model worst-case virus transport scenarios, followed by drywell simulations to investigate the impact of removal parameters on virus transport across various heterogeneity conditions.

All the models discussed in the above paragraph are developed considering unsaturated zone. Since virus contamination highly affects groundwater, the study of virus movement in groundwater is very important. Sim & Chrysikopoulos, (1995)developed a one-dimensional analytical model for virus transport in a homogeneous saturated zone using first-order inactivation rates for both suspended and attached viruses. Accounting for virus sorption and inactivation of a liquid phase and soil-water interface, another model was developed with different time-dependent rate coefficients (Chrysikopoulos & Sim, 1996). One-dimensional models do not account for concentration gradients and contamination transport perpendicular to groundwater flow, whereas multidimensional models do. A three-dimensional model was developed by Sim & Chrysikopoulos, (1998) to determine the movement of viruses in saturated, homogeneous porous media in uniform flow fields. They considered the first-order inactivation of suspended and sorbed viruses with different inactivation rate coefficients. Since studying virus movement in groundwater requires kinetic adsorption of viruses, Schijven et al., (1999) have developed a one-dimensional transport model with one-site kinetic attachments and first-order inactivation rates in the saturated zone. Another model was developed by Schijven et al., (2002) to study the one-dimensional movement of viruses in saturated sand dunes. The model considers two kinetic reversible adsorption sites, and the breakthrough curves obtained from the analysis show better performance than a one-site kinetic model. These

models are designed for homogeneous saturated zones. However, to study the vertical movement of viruses in aquifers, it is important to consider layered geochemical and physical heterogeneity. Thus, Bhattacharjee et al., (2002) developed an effective two-dimensional transport model for predicting virus transport in heterogeneous subsurface porous media. Joshi et al., (2013) utilized the hybrid finite volume method to understand the virus transport process in groundwater by considering both equilibrium and non-equilibrium sorption. They observed that the breakthrough curve in the non-equilibrium model is reported to provide a better fit with the observed data than the equilibrium model. In the public domain, several commercial models (MT3DMS and HYDRUS 3D) can be used to simulate flow and transport (Bedekar et al., 2016; Šimůnek et al., 2012). However, there are some limitations to the models that are currently available. MT3DMS could only simulate flow and transport processes in saturated aquifer flow conditions. Moreover, MT3DMS also consider only equilibrium sorption. HYDRUS-3D do have interfaces for virus transport study for variably saturated zones. However, HYDRUS3D model considers constant attachment-detachment rates for different saturation level. A coupled model for virus and salinity transport in porous media is developed and validated against analytical solutions and experimental data by Zhang et. al (2019). Simulations reveal a previously unaccounted for peak in virus concentration during tidal flow variations in contaminated saline systems, indicating a negative feedback effect not captured by classical models. PFLOTRAN was employed to create a 2D groundwater model investigating seasonal effects on human pathogenic virus transport during induced bank filtration (Knabe et. al., 2023). The results highlighted that diminished travel time played a crucial role in periods with reduced removal of coliforms and somatic coliphages in the aquifer. Rafini et al (2023) develops and validates numerical models for virus transport in saturated porous media. It uses both diagnostic and prognostic approaches to analyze virus concentration dynamics and predict health risks based on dose-response relationships. The models assess the impact of kinetic reaction coefficients and hydraulic parameters on virus transport and compare risk predictions against regulatory standards. The findings aim to inform groundwater regulators, particularly regarding scenarios that may require increased attention due to potential health risks.

2.4 Parameter estimation model:

One must consider the necessary flow and transport parameters to develop an accurate numerical simulation model with proper boundary conditions. Nevertheless, a few parameters cannot be measured on the field or laboratory scale. In such cases, the parameter estimation technique plays a significant role in identifying those parameters. Past studies show that many

parameter estimation models have been developed to estimate the flow and transport parameters. The models available in the literature are briefly discussed in the following section.

2.4.1 Flow parameter estimation:

The modified Richards's equation for unsaturated-saturated flow is solved by adopting numerical methods along with appropriate initial and boundary conditions. In the subsequent step, the numerical model is coupled with the optimization model to estimate soil and hydraulic parameters. The hydraulic parameters are those that define the relationship between hydraulic conductivity (K), volumetric water content (Θ), and pressure head (h). Such parameters are measured or estimated based on different experimental and empirical relations. It is, however, difficult to measure some of these parameters at the desired field or laboratory scale. In practice, if the hydraulic properties of the aquifer are unknown, these must be estimated using hydrogeologic data by the model calibration process. The model calibration process has recently gained significant attention (McLaughlin & Townley, 1996). However, hydraulic parameter identification or inverse problem involves using a mathematical or numerical model to identify hydraulic parameters from field or laboratory observations (Hyun & Lee, 1998). In the subsequent step, the soil and hydraulic parameters are estimated by clubbing the numerical and optimization models. The parameters are estimated by satisfying the objective function, which minimizes the error function between the observed and predicted hydraulic heads(Dane & Hruska, 1983;Kool & Parker, 1987;Yeh, 1986). The observed hydraulic head was obtained from the field study, while the simulated head was obtained by running the numerical simulation model. The optimization model uses various algorithms to provide new solutions to attain the objective function. Estimating parameters in unsaturated flow studies have traditionally been carried out using gradient-based classical optimization methods (Eching et al., 1994;Kool & Parker, 1987; Šimůnek & Van Genuchten, 1996). However, due to the nonlinear behavior of the response function, they sometimes fail to find the optimal global solution to the problem. Woodbury & Rubin, (2000) and Woodbury & Ulrych (2000) applied a full-Bayesian approach using both Bayesian and maximum entropy to estimate transmissivity from the hydrostatic head and transmissivity measurements viewpoints. A simulation-optimization-based model was developed using a meshless local Petrov–Galerkin method and particle swarm algorithm to estimate saturated flow parameters (Swathi & Eldho, 2018).This model predicted only one or two parameters at a time among hydraulic conductivity or transmissivity and specific storage. The model, however, could not provide conclusions about its suitability for different groundwater systems. Another model was developed to estimate the

storage coefficient, transmissivity, and leakage factor by using pumping test data in one-dimensional confined and leaky confined aquifers (Tamer Ayvaz & Gurarslan, 2019). In many groundwater studies, stochastic optimization techniques, such as Pattern Search, Genetic Algorithms, or Simulated Annealing, have been used to reach the optimal global solution. These models were developed to estimate parameters in groundwater aquifers (Huang & Dun, 2008; Şahin, 2018;Samuel & Jha, 2003). A Gaussian process-based iterative EnKF method was applied to estimate the hydraulic parameters for two-dimensional unsaturated flow in layered soil. The model provides improved computational efficiency and comparable results to the standard Iterative Ensemble Kalman Filter (Liu et al., 2020). Another inverse modeling techniques was employed by Coutinho de Oliveira et al (2020) to estimate the soil hydraulic characteristics, including the parameters of the Gardner model for unsaturated soil hydraulic conductivity and the van Genuchten model for soil moisture retention, using field soil moisture measurements obtained during wetting bulb simulation. Later Eryiğit (2021) developed an optimization model for groundwater flow prediction, efficiently estimating multiple parameters with fewer observations and more grid cell numbers compared to previous studies. Implemented with a modified clonal selection algorithm in MATLAB and MODFLOW, the model successfully determined hydraulic conductivity, transmissivity, storage coefficient, and leakance in two- and three-dimensional hypothetical groundwater systems under transient conditions. All such models independently estimated the soil and hydraulic parameters for the unsaturated zone or the saturated zone. Thus, an effective parameter estimation model is yet to develop to estimate the unsaturated and saturated flow parameter together in a single model.

2.4.2 Virus transport parameter estimation:

The linked-simulation optimization model has proven effective in identifying transport parameters in porous media (Abbasi et al., 2003; Bhattacharjya et al., 2015; Jacques et al., 2002; Gribb, 1996; Kodešová et al., 1998, 1999); Inoue et al., 2000; .It is an inverse optimization technique where the objective function minimizes the error between the simulated and observed concentration. However, only a few studies estimated virus transport parameters in the groundwater aquifer. Schijven & Šimůnek, (2002)estimated virus transport parameters in the unsaturated zone using HYDRUS-1D calibrated with measured data iteratively. An estimation of virus transport parameters in saturated zones was also calculated using the Levenberg-Marquard algorithm by Ratha et al., (2009). Although the model could predict two parameters at once, it cannot determine the optimal solution for all four transport parameters simultaneously. As a result, they concluded that virus transport parameters could not be

calculated using classical gradient-based optimization methods. As a solution to such nonlinear nonconvex optimization problems, genetic algorithms (GA) have proved more effective than classical optimization methods when developing various subsurface models (Bhattacharjya & Datta, 2005, 2009; McKinney & Lin, 1994; Morshed & Kaluarachchi, 2000; Rao et al., 2007) Rogers & Dowla, 1994) . Unlike gradient-based algorithms, Genetic algorithm (GA) does not utilize gradient information to obtain optima. The resulting solution may not be the exact global optimal solution but may be close to it. To overcome this, Bhattacharjya et al., (2015) used a hybrid optimization approach to estimate all four parameters simultaneously in a saturated region. The hybrid optimization method starts with a near-global solution obtained by the Genetic algorithm, followed by the simplex method or another local search method to get the solution. Considering equilibrium sorption, Ojha et al., (2012) identified the virus transport parameters in unsaturated zones. There was either an overestimation or an underestimation of the parameters. In addition, only two parameters can be estimated simultaneously. Joshi et al., (2013) utilized the hybrid finite volume model with a nonlinear least squares optimization algorithm to estimate virus transport parameters. Both equilibrium and non-equilibrium sorption theories were considered in developing the model. According to researchers, the breakthrough curve in the non-equilibrium model better fits the observed data than the equilibrium model.

2.5 Source identification model:

Over the past 20 years, several attempts have been made to solve the ADE equation to identify pollution sources. There are three primary methods, including the response matrix method, embedded approach, and linked simulation-optimization approach that provide systematic frameworks for identifying pollution sources in groundwater systems. However, due to non-uniqueness and infinite combinations, one must use the linked simulation-optimization method to get the best solutions. They aim to address the challenge of non-uniqueness and the infinite combinations of potential pollution sources by utilizing optimization techniques to narrow down the possibilities and find the best-fit solutions. Researchers have applied these methods in various studies to successfully identify groundwater pollution sources, as referenced in the works of (Gorelick et al., 1983; Borah & Bhattacharjya, 2013; Leichombam & Bhattacharjya, 2016; Wagner, 1992; Aral & Guan, 1996; Leichombam & Bhattacharjya, 2019;Guneshwor et al., 2018; Rajeev Gandhi et al., 2017; Mahinthakumar & Sayeed, 2005; Xiao et al., 2015;Gurarslan & Karahan, 2015; Yeh et al., 2014).

2.5.1 Linked Simulation-Optimization approach:

The linked simulation-optimization model solves groundwater management problems by externally coupling the simulation and optimization models. This method was developed to overcome the limitations of the response matrix and embedded approach. Any complex simulator can be linked to the optimization model using the linked simulation-optimization model. In this way, it can solve large, heterogeneous, and nonlinear aquifer problems. Groundwater simulators solve the governing equations to simulate the groundwater flow and transport processes. Whenever the optimization model needs information from the groundwater flow and transport simulators, it will repeatedly call the simulation model. To determine the optimal solution iteratively, this simulation model provides the necessary information to the optimization model. Using this one can locate and estimate the contaminant source locations and strength by minimising the absolute difference between observed and simulated concentrations.

Gorelick et al., (1983) tried to solve the source identification problem using a linked simulation-optimization model based on linear programming and a response matrix approach. This model was generated to identify the disposal site for both steady and transient transport cases. Aral et al., (2001) applied linked simulation-optimization to a more complex problem in which the release history and the source location were unknown. It was observed that Progressive Genetic Algorithms (PGA) used in this model is among the most robust algorithms for solving source identification problems compared to other nonlinear issues in this model (Aral & Guan, 1996; Guan & Aral, 1999).Several groundwater management and source identification studies have been performed using different groundwater simulator models, including SUTRA and MODFLOW (Ayvaz, 2010, 2016; Bashi-Azghadi et al., 2016; Bouhlila & Hariga, 2022; Borah & Bhattacharjya, 2013; Datta et al., 2009; Datta et al., 2011; Ngamsritrakul et al., 2015) MT3DMS (Borah & Bhattacharjya, 2016; Leichombam & Bhattacharjya, 2019); FEMWATER(Bhattacharjya et al., 2007; Lin et al., 1997) FEFLOW (C. Zhao et al., 2005; Huo et al., 2007; Matiatos et al., 2019), SEAWAT (Rao, 2006; Kourakos & Mantoglou, 2011, 2013) etc.

The linked simulation-optimization model has proven to have broad applications in groundwater management studies. However, one of the main limitations of the linked model is its computational efficiency. To achieve an optimal solution, the optimization model performs large numbers of simulation calls to simulate the simulation model. Thus, computational

efficiency relies mainly on the simulation model (Bhattacharjya & Datta, 2005). As a result, many researchers attempted to replace this highly complex groundwater simulator with alternative simulators which are computationally less expensive (Hazrati-Yadkoori & Datta, 2017; Zhao et al., 2016; Jiang et al., 2015; Leichombam & Bhattacharjya, 2016; Shahrbanoo Hazrati & Datta, 2017; Sreekanth & Datta, 2011; Xing et al., 2019; Zhao et al., 2020). An alternative groundwater simulator was incorporated using regression models, artificial neural network (ANN), genetic programming, etc. Among all these, ANN has been proven to be one of the best approximate simulators of groundwater that has developed over the years. An ANN model was used to identify source locations of groundwater pollution and determine unknown hydrogeological parameters by (R. M. Singh & Datta, 2004)Based on the model's performance, the model could identify pollution sources and estimate aquifer parameters even when there were measurement errors in the observed concentration data. Bhattacharjya & Datta, (2009) developed an ANN-GA-based simulation-optimization model for controlling saltwater intrusion into the coastal aquifer. According to the evaluated results, the developed model was capable of solving the density-dependent flow and transport processes with the least amount of computation effort. Bhattacharjya et al., (2007) used the FEMWATER program to generate the data to train and validate the ANN model. ANNs have been reported to be highly effective as approximate simulators of density-dependent flow and transport processes in coastal aquifers. According to S. Jiang et al., (2013) a parameter-free harmony search algorithm was employed in the linked simulation optimization model. A satisfactory solution was obtained even for irregular geometries and noisy observations. Datta et al., (2014) developed a software package called GWSID to identify groundwater pollution sources by integrating numerical groundwater flow-transport code with simulated annealing-based optimization model. Borah & Bhattacharjya, (2015) proposed an improved source identification model by incorporating an artificial neural network (ANN) and the groundwater modeling system (GMS) together with the optimization model. The model provides satisfactory results in both computational time and predicting performance. Gurarslan & Karahan, (2015)determined pollution source numbers, locations, fluxes, and release histories using a simulation optimization model. The groundwater flow and transport were simulated using MODFLOW and MT3DMS software. Using accurate and noisy observation data, they developed two hypothetical scenarios. First, the source numbers, location, and active stress period of the pollution sources were assumed to be known, and the release histories were estimated. On the other hand, in the second model, no information about the source is known, and the release histories are determined accordingly. In comparison with previous studies, this result was found to be better.

The model developed so far considers normal solute transport and its source identification in groundwater aquifers. At the same time, there is a need to identify different sources of microbial contamination in groundwater, creating severe health issues for the environment. Rajeev Gandhi et al., (2017) have suggested a virus source identification model when the source number and location are unknown. This is a combination of a discrete and continuous variable problem as the number of sources and their locations are discrete variables, and the source strengths are continuous. The model simulates the virus flow and transport model using MODFLOW and MT3DMS. This simulation model is linked with the optimization model to identify the pollution sources. This model has a few limitations and assumptions; mainly (a) they have used equilibrium sorption to develop the virus transport model. However, one must incorporate kinetic sorption to understand the exact virus transport process in groundwater aquifers. (b) The model considers the source location to be present in the saturated zone. However, virus sources are mainly present in the unsaturated zone, and (c) this model can identify the source location in two-dimensional space. Considering all these limitations, this research tries to develop a virus source identification model by considering the source locations in both saturated and unsaturated zones. The model is also developed incorporating kinetic sorption, and it is formulated to search the location in three-dimensional space.

2.6 Research Gaps:

- Numerous studies have been conducted in recent decades to understand virus transport within porous mediums in saturated and variably saturated environments. In addition, they have also investigated their effects on soil heterogeneities under several other conditions. Although commercial software like MT3DMS and HYDRUS-3D is available in the public domain but they have their own limitations. For example, MT3DMS cannot account for variable degrees of saturation, and HYDRUS 3D cannot handle conditions such as variations in activation and inactivation rates of viruses due to changes in subsurface moisture content, temperatures, etc.
- However, when it comes to developing a coupled unsaturated-saturated groundwater flow model, there is a need for a parameter estimation optimization model that can estimate the unsaturated and saturated flow parameters together. In other words, the existing models are not designed to simultaneously estimate the parameters for both the unsaturated and saturated zones in a coupled manner.

- Traditionally, virus transport models have relied on the constant detachment model to estimate transport parameters. However, this approach may not always be accurate since the detachment-attachment rates of viruses are known to vary based on the saturation level of the medium. Therefore, a more sophisticated approach that accounts for the estimation of variable detachment-attachment rates is necessary.
- In the case of virus source identification, the existing models primarily focus on locating sources in the saturated zone of groundwater. However, in reality, virus sources are often present in the unsaturated zone, which is the region above the water table where the soil or rock is partially saturated with water. By incorporating the unsaturated zone in the virus source identification process, a more comprehensive understanding of the sources and their potential impact on groundwater can be achieved, leading to improved management and remediation strategies.
- Developing a new source identification model that can identify both the source location and source fluxes together, while ensuring efficiency in computational time and predicting performance, is an important task. Researchers can build on the work done by (Borah & Bhattacharjya, 2016) and incorporate additional techniques and algorithms to address the limitation of identifying only the source strength.

CHAPTER 3

THREE-DIMENSIONAL VIRUS TRANSPORT MODEL IN AN UNCONFINED AQUIFER CONSIDERING BOTH THE UNSATURATED AND SATURATED ZONES.

3.1 Overview:

The chapter provides a comprehensive overview of the flow and virus transport models in the unsaturated and saturated zones in groundwater aquifers. The chapter has four major sections:

- Introduction: This section provides a brief overview of the significance of the study on the virus transport movement.
- Methodology: This section will describe the development of the flow and virus transport model, including the numerical methods used, the assumptions and simplifications made, and the input data requirements. It would also cover the calibration and validation of the model, including the model performance evaluation using statistical metrics.
- Results and Discussion: This section will present the flow and virus transport modeling results, including the spatial and temporal distribution of the virus concentration. It would also discuss the findings and their implications for virus source identification and risk assessment.
- Conclusion: This section will summarize the study's key findings, highlight the modeling approach's contributions and limitations, and provide recommendations for future research.

Overall, the chapter will give a detailed account of the modeling approach and methodology used to simulate the flow and virus transport in unsaturated and saturated zones of a groundwater aquifer and offer insights into the fate and transport of viruses in the subsurface environment.

3.2 Introduction:

Groundwater is an essential source of freshwater for human consumption, irrigation, and industrial purposes. However, anthropogenic activities such as urbanization, industrialization, and agricultural practices have led to the contamination of groundwater resources. Pollutants such as heavy metals, pesticides, fertilizers, and organic compounds can seep through the soil

and enter the groundwater system, compromising the quality and safety of the water. In addition, inadequate sanitation practices also lead to the introduction of viruses and other pathogens into the groundwater system, posing a significant health risk to the population. As such, effective management and protection of groundwater resources are essential to ensure the availability of safe and clean water for future generations. Understanding how these microbes enter the subsurface is crucial because these sources are often found in unsaturated zones. A significant portion of the precipitation that falls onto the earth's surface enters the sub-surface zone by infiltration. The unsaturated zone sometimes acts as a filter, removing some pollutants and pathogens, but it is not always effective in preventing their transport into the saturated groundwater aquifer beneath. Pathogens, including viruses, can travel through the unsaturated zone and enter the groundwater system, where they can persist for extended periods. Developing models that simulate virus transport in unsaturated and saturated zones can help identify potential sources of contamination, predict the spread of viruses, and design effective management strategies to protect groundwater resources.

The flow and transport processes can be mathematically modeled through partial differential equations using appropriate initial and boundary conditions. Modeling virus transport in the unsaturated and saturated zone is complex as the moisture velocity within these zones depends significantly on the degree of saturation. When this situation occurs, it becomes necessary to solve the flow equation before solving the virus transport equation. Using this model, one can conceptualize hydro-geological processes and systematically integrate field data. Numerical methods such as finite elements, finite differences, and characteristic methods can be used to model groundwater flow. The results obtained from the flow model will be considered in the virus transport equation to study the movement of the virus from the source to the groundwater aquifer. Several virus transport models have been proposed so far, but each model has its own limitations and may be ineffective. As mentioned by different researchers virus inactivation and sorption rates depend on the degree of saturation (Chu et al., 2001;Jin et al., 2000;Lance and Gerba, 1984;Powelson et al.,1990;Poletika et al., 1995). However, one way to model the virus transport process with a constant sorption rate is by assigning a more significant yet constant value of attachment-detachment rates. Such a model has been referred to as the ‘constant-detachment model’(Torkzaban et al. 2006a; 2006b). This dependency of the attachment and detachment rate coefficient on fluid saturation should be considered. In this study, an alternative and somewhat simpler formulation are presented by assuming variable attachment and detachment rates of virus for different saturation levels, and thus termed as ‘variable detachment model’.

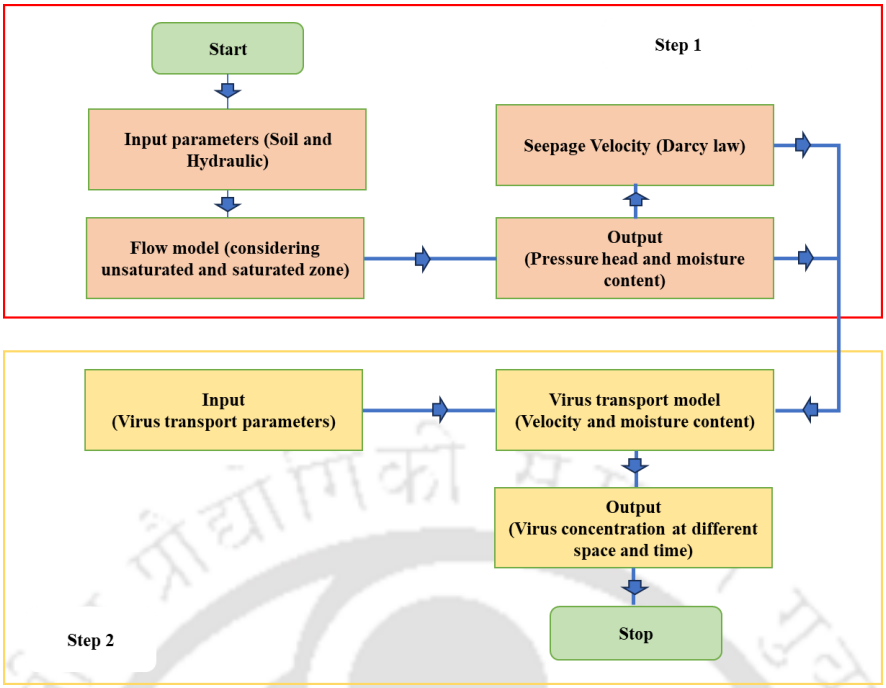


Figure 3. 1: Systematic representation of the coupled Flow and transport model in an unconfined aquifer.

The model is developed using a Block-centered finite difference scheme for spatial variation and a backward finite difference scheme for temporal variation. The model undergoes two steps; as a first step, the model solves the three-dimensional modified form of the Richards equation continuously across the entire flow domain (from the ground surface to the impervious media at the bottom). This equation is highly nonlinear, so to remove the equation's nonlinearity, iteration methods will be needed to remove the nonlinearity. The modified Picard iterative method is adopted in this study to solve this problem. Using the flow model, the solution will provide the pressure head and moisture content at each node at the specified time step. Later, by using Darcy's law, the seepage velocity is calculated based on the pressure head of the solution. In the second step, the three-dimensional virus transport equation is solved, which describes the mechanism of advection, dispersion, adsorption, and inactivation within the transport process. By incorporating the obtained seepage velocity and moisture content into the virus transport equation, the solution to the virus transport equation is obtained. The model will provide breakthrough curves of the viruses at different observation locations over time. The methodology proposed to develop the virus transport model is shown in Figure 3. 1.

3.3 Methodology of Flow model:

The Richards equation is a partial differential equation commonly used to describe the flow of water through unsaturated porous media, such as soil. It has three formulations: pressure-based,

moisture-based, and mixed-based. Pressure-based uses pressure head as the primary unknown, moisture-based uses moisture content, while mixed incorporates both. Moreover, the mixed form of the Richards equation is computationally efficient and can model various problems, including infiltration into extremely dry soils (Celia et al., 1990 and Clement et al., 1994). Dogan & Motz, (2005) modified the mixed form of the Richards equation by incorporating both saturated and unsaturated zones in the flow domain, which is discussed in this section.

For the unsaturated zone the three-dimensional mixed form of the Richards' equation is given by equation (3.1):

$$\frac{\partial}{\partial x}\left(K_{xx}(h)\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}(h)\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}(h)\frac{\partial h}{\partial z}+K_{zz}(h)\right)+q_e=\frac{\partial \theta}{\partial t} \tag{3.1}$$

Where, θ is the water content, h is the pressure head, K_{xx}, K_{yy}, K_{zz} are the hydraulic conductivity along x,y,z direction considering the coordinate system as the principal directions of the hydraulic conductivity tensor. qe is the source or sink term.

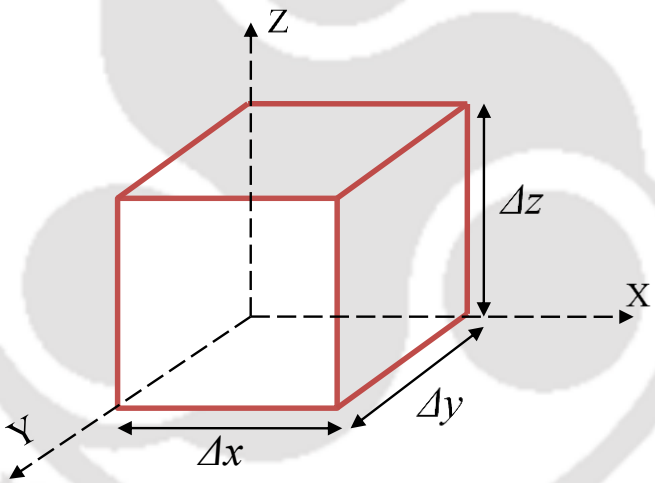


Figure 3.2: 3D volume showing unit vector outward.

Integrating this equation (3.1) for the whole control volume as shown in Figure 3.2,

$$\iiint_V \frac{\partial}{\partial x}\left(K_{xx}(h)\frac{\partial h}{\partial x}\right)dV+\iiint_V \frac{\partial}{\partial y}\left(K_{yy}(h)\frac{\partial h}{\partial y}\right)dV+\iiint_V \frac{\partial}{\partial z}\left(K_{zz}(h)\frac{\partial h}{\partial z}+K_{zz}(h)\right)dV$$
$$+\iiint_V q_e dV=\iiint_V \frac{\partial \theta}{\partial t}dV \tag{3.2}$$

Using the Gauss divergence theorem

$$\iiint_V (\nabla . F)dV=\iint_S F.\hat{n}\,ds \tag{3.3}$$

Now using equation (3.3) in equation (3.2) we have,

$$\left(K_{xx}(h)\frac{\partial h}{\partial x}\right)(\Delta y\Delta z)-\left(K_{xx}(h)\frac{\partial h}{\partial x}\right)(\Delta y\Delta z)+\left(K_{yy}(h)\frac{\partial h}{\partial y}\right)(\Delta x\Delta z)-\left(K_{yy}(h)\frac{\partial h}{\partial y}\right)(\Delta x\Delta z)$$
$$+\left(K_{zz}(h)\frac{\partial h}{\partial z}+K_{zz}(h)\right)(\Delta x\Delta y)-\left(K_{zz}(h)\frac{\partial h}{\partial z}+K_{zz}(h)\right)(\Delta x\Delta y)+q_e(\Delta x\Delta y\Delta z)=\frac{\partial\theta}{\partial t}(\Delta x\Delta y\Delta z) \quad (3.4)$$

Thus, the continuity equation to a unit volume of aquifer in unsaturated zone is given by equation (3.5)

$$\sum Q_I - Q_O \pm Q_e = \frac{\partial\theta}{\partial t}\Delta V \quad (3.5)$$

Where, Q_I is the inflow rate into the cell [L^3T^{-1}], Q_o is the outflow rate from the cell [L^3T^{-1}], Q_e represents pumping or recharge rate [L^3T^{-1}]. ΔV is the volume of the cell [L^3].

For Saturated zone the three-dimensional flow equation is given by equation (3.6):

$$K_{xx}\frac{\partial}{\partial x}\left(\frac{\partial H}{\partial x}\right)+K_{yy}\frac{\partial}{\partial y}\left(\frac{\partial H}{\partial y}\right)+K_{zz}\frac{\partial}{\partial z}\left(\frac{\partial H}{\partial z}\right)+q_e=S_wS_s\frac{\partial H}{\partial t} \quad (3.6)$$

Where, S_w is the degree of saturation (For saturated zone S_w=1and for unsaturated zone S_w<1), S_s is the specific storage and H is the hydraulic head obtained by summation of pressure head(h) and datum head(z).

Integrating this equation for the whole control volume,

$$\iiint_V K_{xx}\frac{\partial}{\partial x}\left(\frac{\partial H}{\partial x}\right)dV+\iiint_V K_{yy}\frac{\partial}{\partial y}\left(\frac{\partial H}{\partial y}\right)dV+\iiint_V K_{zz}\frac{\partial}{\partial z}\left(\frac{\partial H}{\partial z}\right)dV+\iiint_V q_e dV$$
$$=\iiint_V S_wS_s\frac{\partial H}{\partial t}dV \quad (3.7)$$

Applying Gauss divergence theorem, we can express equation (3.7) as follows:

$$K_{xx}\left(\frac{\partial H}{\partial x}\right)(\Delta y\Delta z)-K_{xx}\left(\frac{\partial H}{\partial x}\right)(\Delta y\Delta z)+K_{yy}\left(\frac{\partial H}{\partial y}\right)(\Delta x\Delta z)$$
$$-K_{yy}\left(\frac{\partial H}{\partial y}\right)(\Delta x\Delta z)+K_{zz}\left(\frac{\partial H}{\partial z}\right)(\Delta x\Delta y)-K_{zz}\left(\frac{\partial H}{\partial z}\right)(\Delta x\Delta y)+q_e(\Delta x\Delta y\Delta z) \quad (3.8)$$
$$=S_wS_s\frac{\partial H}{\partial t}(\Delta x\Delta y\Delta z)$$

Thus, the continuity equation to a unit volume of aquifer in saturated zone is given by equation (3.5)

$$\sum Q_I - Q_O \pm Q_e = S_wS_s\frac{\Delta H}{\Delta t}\Delta V \quad (3.9)$$

3.3.1 Governing equation for flow model:

Considering the unsaturated and saturated flow condition in the porous media as given by equation (3.5) and (3.9) Dogan & Motz, (2005) has provided the modified form of Richard equation as

$$\frac{\partial}{\partial x}\left(K_{xx}(h)\frac{\partial h}{\partial x}\right)+\frac{\partial}{\partial y}\left(K_{yy}(h)\frac{\partial h}{\partial y}\right)+\frac{\partial}{\partial z}\left(K_{zz}(h)\frac{\partial h}{\partial z}+K_{zz}(h)\right)+q_e=\frac{\partial \theta}{\partial t}+S_wS_s\frac{\partial h}{\partial t} \quad (3.10)$$

Considering the total hydraulic head (H), the equation (3.10) can be rewritten as follows

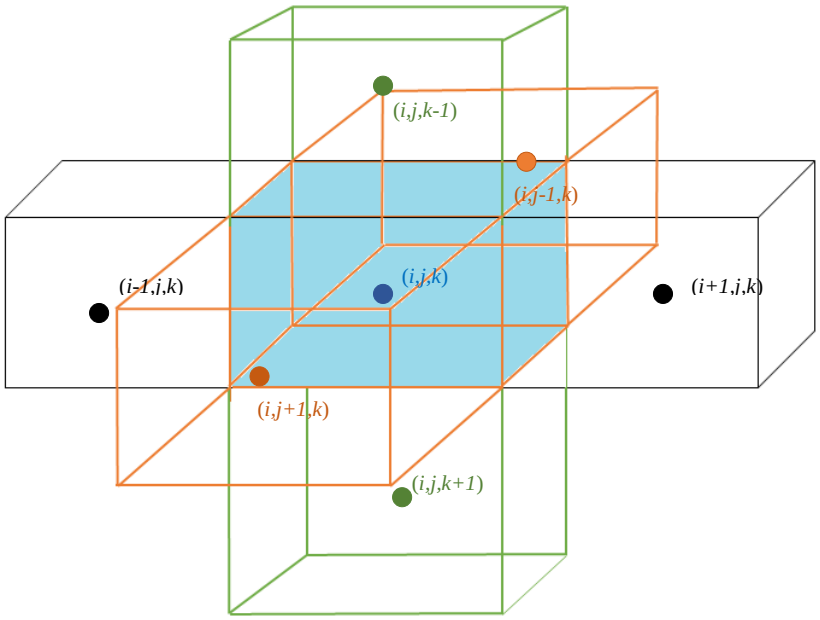
$$K_{xx}\frac{\partial}{\partial x}\left(\frac{\partial H}{\partial x}\right)+K_{yy}\frac{\partial}{\partial y}\left(\frac{\partial H}{\partial y}\right)+K_{zz}\frac{\partial}{\partial z}\left(\frac{\partial H}{\partial z}\right)+q_e=\frac{\partial \theta}{\partial t}+S_wS_s\frac{\partial H}{\partial t} \quad (3.11)$$

The equation (3.11) is the modified form of the Richards equation and is solved in this thesis using numerical approximation.

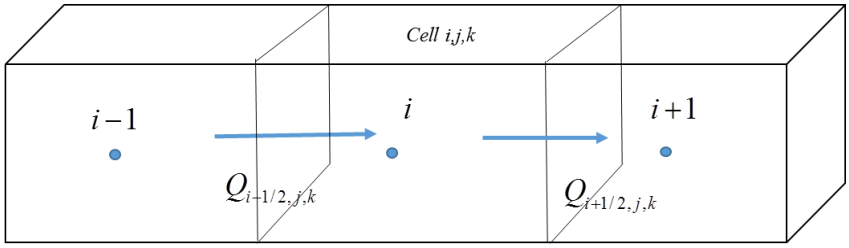
3.3.2 Numerical approximation used in flow model:

This study uses the implicit block-centered finite difference method to develop the numerical groundwater flow model. To develop the model, the sum of water flowing into and out of a unit volume of cell in an aquifer must be equal to the rate of change in the volume of storage within the cell. Based on the assumption of constant groundwater density, the flow model is given by:

$$\sum Q_I - Q_O + Q_e = \left(\frac{\partial \theta(h)}{\partial t}+S_wS_s\frac{\partial H}{\partial t}\right)\Delta V \quad (3.12)$$



(a) The finite difference grid system used in three- dimensional space. The interfaces between node (i,j,k) and the adjacent six nodes are shown in blue colour.



(b) Inflow and outflow from cell (*ijk*) along X-axis

Figure 3.3: Graphical representation of the inflow and outflow of a cell (*ijk*) along x, y, and z-axis.

Figure 3.3 shows the finite difference of a three-dimensional space, where the subscripts *i,j,k* represents the *x,y* and *z*-direction. The inflow and outflow along the six faces of the cell (*ijk*) along x, y, and z-axis is shown. The inflow and outflow taking place in the cell (*i,j,k*) along x-axis can be expressed using the Darcy Buckingham equation. The flux coming into the cell (*i,j,k*) along x-axis is given by equation (3.13a) and the flux coming out from the same cell given by equation (3.13b).

$$Q_{i-1/2,j,k} = CN_{i-1/2,j,k}\left(\Delta y_j \Delta z_k\right)\left(H_{i,j,k} - H_{i-1,j,k}\right) \tag{3.13a}$$

$$Q_{i+1/2,j,k} = CN_{i+1/2,j,k}\left(\Delta y_j \Delta z_k\right)\left(H_{i+1,j,k} - H_{i,j,k}\right) \tag{3.13b}$$

Similarly, for y-axis it is given as follows

$$Q_{i,j-1/2,k} = CN_{i,j-1/2,k}\left(\Delta x_i \Delta z_k\right)\left(H_{i,j,k} - H_{i,j-1,,k}\right) \tag{3.14a}$$

$$Q_{i,j+1/2,k} = CN_{i,j+1/2,k}\left(\Delta x_i \Delta z_k\right)\left(H_{i,j+1,k} - H_{i,j,,k}\right) \tag{3.14b}$$

For z-axis it is given as follows

$$Q_{i,j,k-1/2} = CN_{i,j,k-1/2}\left(\Delta x_i \Delta y_j\right)\left(H_{i,j,k} - H_{i,j,k-1}\right) \tag{3.15a}$$

$$Q_{i,j,k+1/2} = CN_{i,j,k+1/2}\left(\Delta x_i \Delta y_j\right)\left(H_{i,j,k+1} - H_{i,j,k}\right) \tag{3.15b}$$

Where, CN represents the conductance between neighboring nodes, that is, $CN_{i-1/2,j,k}$ is the conductance between node (*i–1,j,k*) and node (*i,j,k*). The details about the conductance term is given by the equation (3.16) as given by Dogan et al., (2005).

$$CN_{i+1/2,j,k} = -\frac{\left(K_S K_r\right)_{i+1/2,j,k}}{\left(\Delta x_i + \Delta x_{i+1}\right)/2} \tag{3.16a}$$

$$CN_{i-1/2,j,k} = -\frac{\left(K_S K_r\right)_{i-1/2,j,k}}{\left(\Delta x_i + \Delta x_{i-1}\right)/2} \tag{3.16b}$$

Where, $K_r$is the ratio of the unsaturated hydraulic conductivity to the saturated hydraulic conductivity ($K(h)/K_s$). K_r is equal to 1.0 for fully saturated condition and 0.0 for very dry unsaturated condition. The average of the hydraulic conductivity is discussed in detailed in section (3.2.3).

Now, putting the expression for inflow and outflow flux with respect to the cell (*ijk)*along x, y, and z direction in the equation (3.12) we have,

$$\begin{aligned} &CN_{i-1/2,j,k}\left(\Delta y_j \Delta z_k\right)\left(H_{i,j,k} - H_{i-1,j,k}\right) - CN_{i+1/2,j,k}\left(\Delta y_j \Delta z_k\right)\left(H_{i+1,j,k} - H_{i,j,k}\right) + CN_{i,j-1/2,k}\left(\Delta x_i \Delta z_k\right) \\ &\left(H_{i,j,k} - H_{i,j-1,k}\right) - CN_{i,j+1/2,k}\left(\Delta x_i \Delta z_k\right)\left(H_{i,j+1,k} - H_{i,j,k}\right) + CN_{i,j,k-1/2}\left(\Delta x_i \Delta y_j\right)\left(H_{i,j,k} - H_{i,j,k-1}\right) \\ &- CN_{i,j,k+1/2}\left(\Delta x_i \Delta y_j\right)\left(H_{i,j,k+1} - H_{i,j,k}\right) = \left(\frac{\partial \theta(h)}{\partial t} + S_w S_s \frac{\partial H}{\partial t}\right)\left(\Delta x_i \Delta y_j \Delta z_k\right) \end{aligned} \tag{3.17}$$

$$\begin{aligned} &CN^{t+1,m}_{i-1/2,j,k}\left(\frac{H^{t+1,m+1}_{i,j,k} - H^{t+1,m+1}_{i-1,j,k}}{\Delta x_i}\right) - CN^{t+1,m}_{i+1/2,j,k}\left(\frac{H^{t+1,m+1}_{i+1,j,k} - H^{t+1,m+1}_{i,j,k}}{\Delta x_i}\right) + CN^{t+1,m}_{i,j-1/2,k}\left(\frac{H^{t+1,m+1}_{i,j,k} - H^{t+1,m+1}_{i,j-1,k}}{\Delta y_j}\right) \\ &- CN^{t+1,m}_{i,j+1/2,k}\left(\frac{H^{t+1,m+1}_{i,j+1,k} - H^{t+1,m+1}_{i,j,k}}{\Delta y_j}\right) + CN^{t+1,m}_{i,j,k-1/2}\left(\frac{H^{t+1,m+1}_{i,j,k} - H^{t+1,m+1}_{i,j,k-1}}{\Delta z_k}\right) - CN^{t+1,m}_{i,j,k+1/2}\left(\frac{H^{t+1,m+1}_{i,j,k+1} - H^{t+1,m+1}_{i,j,k}}{\Delta z_k}\right) \\ &+ \frac{Q_e^{t+1,m}}{\Delta x_i \Delta y_j \Delta z_k} = \left(\frac{\theta(h^{t+1,m+1}_{i,j,k}) - \theta(h^{t}_{i,j,k})}{t^{t+1} - t^{t}}\right) + S_w S_s\left(\frac{H^{t+1,m+1}_{i,j,k} - H^{t}_{i,j,k}}{t^{t+1} - t^{t}}\right) \end{aligned} \tag{3.18}$$

Where, ‘t’ represents the previous time step, and ‘t+1’ represents the current time step. The expression *Θ(h)* and *K(h)* are highly non-linear, making the governing equation complex. Therefore, iterative methods are necessary to overcome the non-linearity. This study uses a backward Euler approximation with the Picard iteration to simplify time discretization. The Picard iteration is denoted by ‘m’ for the first iteration and ‘m+1’ for the next.

Since the value of θ at m+1 iteration is unknown so the value of $\theta^{t+1,m+1}_{i,j,k}$ can be expressed using the Taylor series as given by (Dogan et al., (2005)

$$\theta^{t+1,m+1}_{i,j,k} = \theta^{t+1,m}_{i,j,k} + \left(\frac{d\theta}{dh}\right)^{t+1,m}_{i,j,k}\left(H^{t+1,m+1}_{i,j,k} - H^{t+1,m}_{i,j,k}\right) \tag{3.19}$$

Putting the value of $\theta^{t+1,m+1}_{i,j,k}$ in the equation (3.18) we have

$$CN^{t+1,m}_{i-1/2,j,k}\left(\frac{H^{t+1,m+1}_{i,j,k}-H^{t+1,m+1}_{i-1,j,k}}{\Delta x_i}\right)-CN^{t+1,m}_{i+1/2,j,k}\left(\frac{H^{t+1,m+1}_{i+1,j,k}-H^{t+1,m+1}_{i,j,k}}{\Delta x_i}\right)+CN^{t+1,m}_{i,j-1/2,k}\left(\frac{H^{t+1,m+1}_{i,j,k}-H^{t+1,m+1}_{i,j-1,k}}{\Delta y_j}\right)$$
$$-CN^{t+1,m}_{i,j+1/2,k}\left(\frac{H^{t+1,m+1}_{i,j+1,k}-H^{t+1,m+1}_{i,j,k}}{\Delta y_j}\right)+CN^{t+1,m}_{i,j,k-1/2}\left(\frac{H^{t+1,m+1}_{i,j,k}-H^{t+1,m+1}_{i,j,k-1}}{\Delta z_k}\right)-CN^{t+1,m}_{i,j,k+1/2}\left(\frac{H^{t+1,m+1}_{i,j,k+1}-H^{t+1,m+1}_{i,j,k}}{\Delta z_k}\right)$$
$$+\frac{Q_e{}^{t+1,m}}{\Delta x_i\Delta y_j\Delta z_k}=\left(\frac{\theta(h)^{t+1,m}_{i,j,k}-\theta(h)^{t}_{i,j,k}}{t^{t+1}-t^{t}}\right)+C(h)^{t+1,k}_{i,j,k}\left(\frac{H^{t+1,m+1}_{i,j,k}-H^{t+1,m}_{i,j,k}}{t^{t+1}-t^{t}}\right)+S_wS_s\left(\frac{H^{t+1,m+1}_{i,j,k}-H^{t}_{i,j,k}}{t^{t+1}-t^{t}}\right)$$

(3.20)

From the previous equation, the finite- difference expression for the spatial and temporal derivatives can be rearranged by coining the unknown terms on the left and the known terms on the right of the equation. The rewritten equation is given by (3.21).

$$aH^{t+1,m+1}_{i-1,j,k}+bH^{t+1,m+1}_{i,j-1,k}+cH^{t+1,m+1}_{i,j,k-1}+dH^{t+1,m+1}_{i,j,k}+eH^{t+1,m+1}_{i+1,j,k}+fH^{t+1,m+1}_{i,j+1,k}+gH^{t+1,m+1}_{i,j,k+1}$$
$$=s-a1.H^{t+1,m}_{i,j,k}-a2.H^{t}_{i,j,k}-\frac{Q_e{}^{t+1,m}}{\Delta x_i\Delta y_j\Delta z_k} \qquad (3.21)$$

Where,

$$a=-\frac{CN^{t+1,m}_{i-1/2,j,k}}{\Delta x_i}$$

$$b=-\frac{CN^{t+1,m}_{i,j-1/2,k}}{\Delta y_j}$$

$$c=-\frac{CN^{t+1,m}_{i,j,k-1/2}}{\Delta z_k}$$

$$e=-\frac{CN^{t+1,m}_{i+1/2,j,k}}{\Delta x_i}$$

$$f=-\frac{CN^{t+1,m}_{i,j+1/2,k}}{\Delta y_j}$$

$$g=-\frac{CN^{t+1,m}_{i,j,k+1/2}}{\Delta z_k}$$

$$s=\frac{\theta^{t+1,m}_{i,j,k}-\theta^{t}_{i,j,k}}{t^{t+1}-t^{t}}$$

$$r1=\frac{C(h)^{t+1,m}_{i,j,k}}{t^{t+1}-t^{t}}$$

$$r2 = \frac{S_w S_s}{t^{t+1} - t^t}$$

$$d = -(a + b + c + e + f + g + r1 + r2)$$

The equation (3.21) is the final form of the finite difference equation of one cell. This equation is solved at all the nodes of the interior cell of the flow domain.

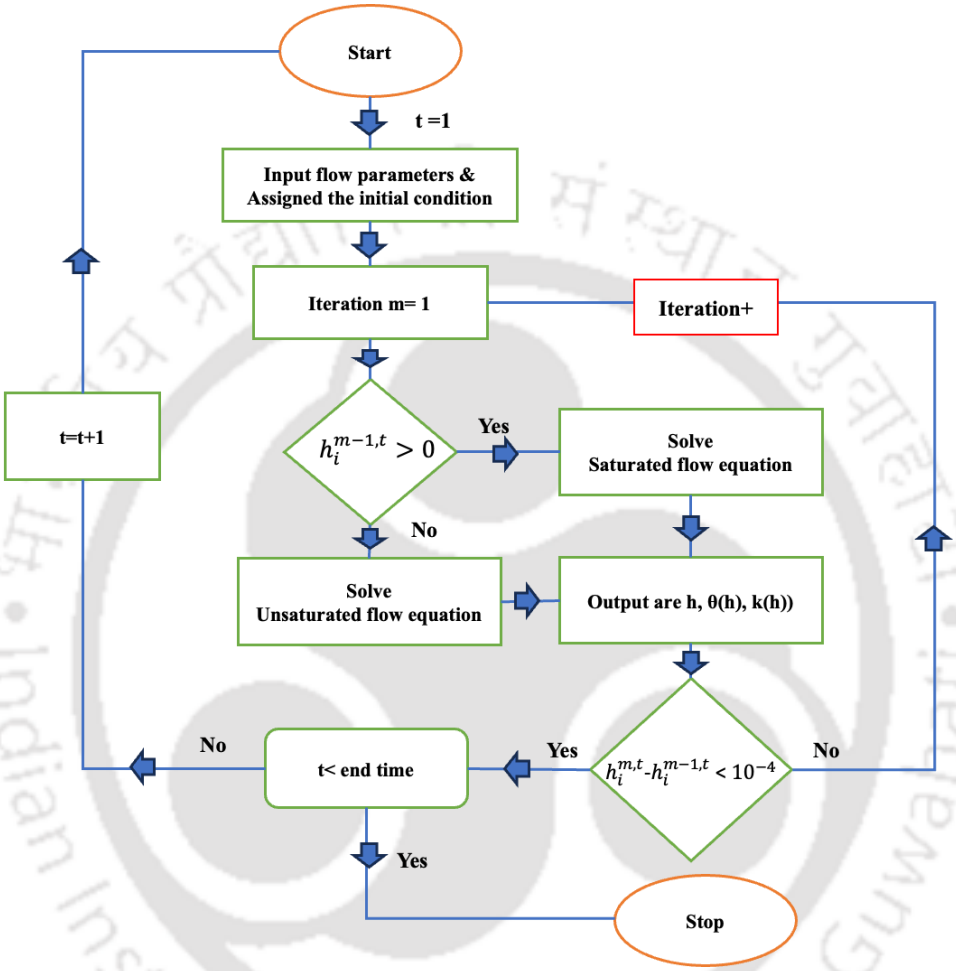


Figure 3. 4: Flowchart of the step by step procedure for flow model.

An appropriate boundary condition is incorporated at boundary nodes in this equation. Thus, the resulting linear equation in matrix form is given by equation (3.22)

$$[A]^{t+1,m}[H]^{t+1,m+1} = [RHS]^{t+1,m} \tag{3.22}$$

where [A] is a square matrix consisting of the coefficients(a,b,c,d,e,f,g) of the finite difference equation(3.22), {H} is unknown hydraulic head values for current time step, and {RHS} is the forcing vector consisting of known values from the previous time step or previous iteration values. This linear equation is solved at every grid by removing the nonlinear terms in the matrix [A] at every modified Picard iteration level. This problem is solved using the preconditioned conjugate gradient method (PCGM), which is more memory-efficient than

other iterative methods and has a faster convergence rate (Celia et al., 1990; Clement et al., 1994; Dogan & Motz, 2005). This three-dimensional groundwater flow model is developed in MATLAB (2017) environment. The step by step procedure to develop the flow model is given in Figure 3. 4.

3.3.3 Averaging Hydraulic conductivity:

According to (Lappala et al., 1987), distance-weighted harmonic means perform better in saturated soil than other methods, while distance-weighted arithmetic means perform better in unsaturated soil. As Darcy's equation illustrates, water will flow from one cell containing moisture to another that is dry in the unsaturated zone. In such cases, if the harmonic mean method were used, the average hydraulic conductivity between these two cells would be equal to zero. As a result, water will not flow from the wet cell to the dry cell, thus, contradicting Darcy's equation. This study uses weighted arithmetic means to calculate the mean hydraulic conductivity between cells that are anticipated to be unsaturated using equation (3.23a). The distance-weighted harmonic mean is used for saturated cells according to equation (3.23b).

For unsaturated zone, distance- weighted arithmetic mean is used as given by Dogan & Motz, (2005) and is expressed as:

$$\left(K_S K_r\right)_{i-1/2,j,k} = \frac{\Delta x_i \left(K_S K_r\right)_{i,j,k} + \Delta x_i \left(K_S K_r\right)_{i-1/2,j,k}}{\left(\Delta x_i + \Delta x_{i-1}\right)} \tag{3.23a}$$

For saturated zone, distance- weighted harmonic mean is used as given by Dogan & Motz, (2005) and is expressed as:

$$\left(K_S\right)_{i-1/2,j,k} = \frac{\left(\Delta x_i + \Delta x_{i-1}\right)}{\frac{\Delta x_i}{\left(K_S\right)_{i,j,k}} + \frac{\Delta x_{i-1}}{\left(K_S\right)_{i-1/2,j,k}}} \tag{3.23b}$$

3.3.4 Boundary condition:

Boundary conditions can be separated into three types: Neumann flux (specified flux), Drichlet pressure head (specified pressure head), and a combination of Neumann and Drichlet conditions (as variable condition).The prescribed values for each of these boundary conditions can either be constant or variable over time. Rainfall, infiltration, and evaporation processes can be described using flux boundary conditions. In the case of a lake or a river with a constant head, specified head boundary conditions can be defined. Rainfall and evaporation of soil from the surface are described by variable boundary conditions. In this study Neumann and Drichlet boundary condition is adopted.

The specified pressure head (Dirichlet) boundary condition is given as follows

$$H_{i,j,k} = \left(H_d\right)_{i,j,k} \tag{3.24a}$$

We are considering two conditions for the Neumann boundary condition: one for no flow condition and another for a constant flux. The boundary condition is expressed by equation (3.24b) and (3.24d).

For No-flow condition:

$$H_{N,j,k} = H_{N+1,j,k} \tag{3.24b}$$

For Constant flux boundary:

$$Q_N = Q_{i,j,N+1/2} = CN_{i,j,N+1/2}\left(\Delta x_i \Delta y_j\right)\left(H_{i,j,N+1} - H_{i,j,N}\right) \tag{3.24c}$$

Where, $H_{N+1,j,k}$ represents the imaginary node outside the boundary along x-direction. H_d is the specified hydraulic head at the boundary node, Q_N is the specified flux in the z-direction at the node (i,j,N+1/2); N+1/2 is the top boundary in the z-direction.

3.3.5 Constitutive relations:

The modified form of the Richards equation given by equation (3.18) shows that the Hydraulic conductivity *K(h)* and specific moisture content *C(h)* are both dependent on pressure head, which makes the equation highly nonlinear. In order to linearize *K(h)* and *C(h)*, a relationship between pressure head (*h*) and moisture content (Θ) is necessary at every level of the Picard algorithm. As such several empirical equations were developed based on the moisture retention curve that calculates hydraulic conductivity as a function of pressure head (Brooks & Corey, 1964; van Genuchten, 1980; Van Genuchten, M. Th. & Nielsen, 1985). As part of this study, the flow model is developed using Van Genuchten, M. Th. & Nielsen, (1985).

Constitutive relation for K(h):

For $h < 0$

$$K_r = \frac{K(h)}{K_s} = \left(1+\beta\right)^{-\frac{5}{2}(1-1/n)}\left[\left(1+\beta\right)^{(1-1/n)} - \beta^{(1-1/n)}\right]^2 \tag{3.25a}$$

For $h \geq 0$

$$K_r = \frac{K(h)}{K_s} = 1 \tag{3.25b}$$

Constitutive relation for C(h):

When $h \leq h_0$

$$C(h) = \frac{(n-1)\left(\theta_s - \theta_r\right)|h|^{n-1}}{|h_a|^n\left(1+\beta\right)^{2-1/n}} \tag{3.25c}$$

When $h > h_0$

$$C(h) = 0 \tag{3.25d}$$

Constitutive relation for θ (h):

When $h \le h_0$ $$\theta(h) = \theta_r + (\theta_s - \theta_r)(1+\beta)^{(1/n-1)} \quad (3.25e)$$

When $h > h_0$ $$\theta(h) = \theta_r + (\theta_s - \theta_r)(1+\beta_0)^{1/n-1} + S_S(h - h_0) \quad (3.25f)$$

where,$\beta = \left|\frac{h}{h_a}\right|^n$,h_a is the air entry pressure [L] and reciprocal of soil parameter (α), n is the fitting parameter in the moisture retention curve,$\beta_0 = \left|\frac{h_0}{h_a}\right|^n$, and h_0 is a parameter depending upon the Specific storage (S_S).

$$S_s = \left.\frac{(n-1)(\theta_s - \theta_r)|h|^{n-1}}{|h_a|^n(1+\beta)^{2-1/n}}\right|_{h=h_0} \quad (3.26)$$

When, $h \ge h_0$, the equation (3.25) solves for the saturated flow condition, i.e., $C(h) = 0$, $K(h) = K_s$, and $S_w = 1$. When $h < h_0$, then the equation (3.25) solves for the unsaturated flow condition, where, $C(h) \neq 0$, $K(h)$ is the function of pressure head, $S_w < 1$, and $S_s = 0$. By solving the three-dimensional groundwater flow equation using numerical methods, the hydraulic head distribution, gradient and moisture content can be obtained at each grid cell of the model domain. Using these values, the seepage velocity is calculated for each interfaces of the grid cell in the model domain, which is then used in the virus transport model to simulate the movement of viruses through the aquifer.

3.3.6 Calculation of the velocity flux at the cell interfaces:

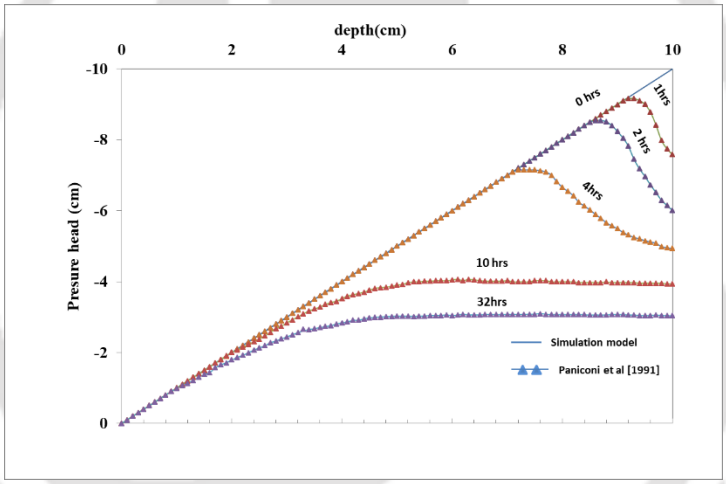
Using the hydraulic head, the seepage velocity is calculated for each interfaces of the grid cell in the model domain is calculated using Darcy law. Darcy's law is a fundamental principle in fluid mechanics and is used to describe the flow of fluids (e.g., water) through porous media. The velocity flux at the interfaces along the x-direction is calculated using the Darcy law as follow:

$$q^{t+1}_{i-1/2,j,k} = -CN_{i-1,j,k}\left(H^{t+1}_{i,j,k} - H^{t+1}_{i-1,j,k}\right)(\Delta y \Delta z) \quad (3.27)$$

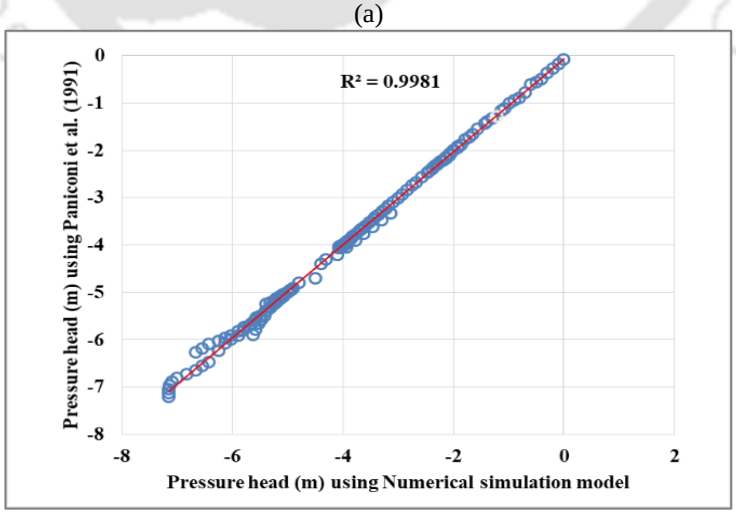
The negative sign in the equation indicates that the flow occurs in the direction of decreasing hydraulic head. Similarly, the velocity flux along y and z axis is calculated using the equation (3.27). This velocity flux is incorporated in the virus transport model to simulate the movement of viruses through the aquifer.

3.3.7 Validation of the saturated-unsaturated flow model:

To evaluate the accuracy of the numerical solution of the saturated-unsaturated flow model, two example problems from the literature were selected. Firstly, a one-dimensional flow model with transient infiltration toward the water table is selected (Paniconi et al., 1991). The developed three-dimensional model is applied to a reduced one-dimensional problem by keeping the other two dimensions (x and y) equal to 1. The model is then simulated for 32 hours with the same boundary conditions and input parameters as given by Paniconi et al., (1991).Figure 3. 5 shows the comparison of the solutions obtained from the numerical simulation model that was developed using the code written in MATLAB in this research with the solution obtained by Paniconi et al., (1991). The scatter plots Figure 3. 5(b) correlates the pressure head (m) obtained from Paniconi et al., (1991) and Numerical simulation model. The regression coefficient is 0.9981, which ascertains that the numerical simulation groundwater flow model could provide an accurate solution as obtained by Paniconi et al (1991).



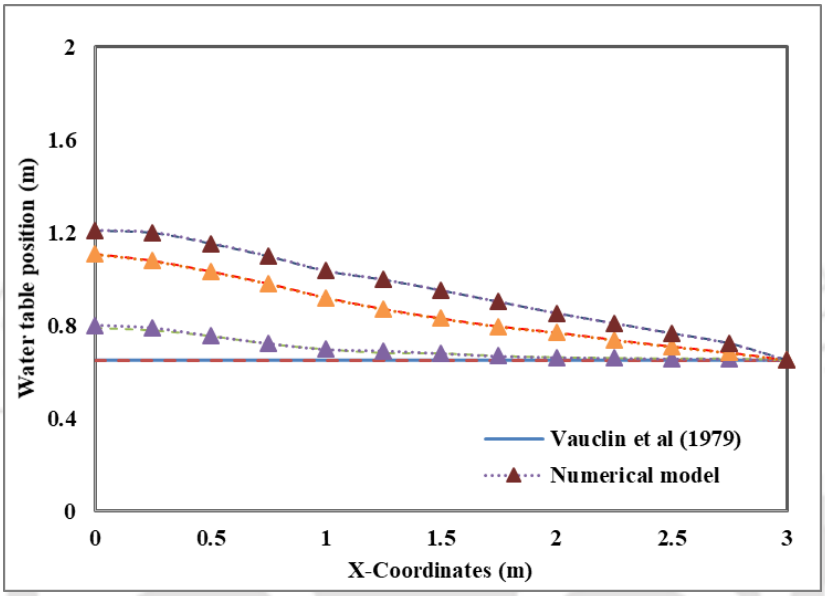
(a)



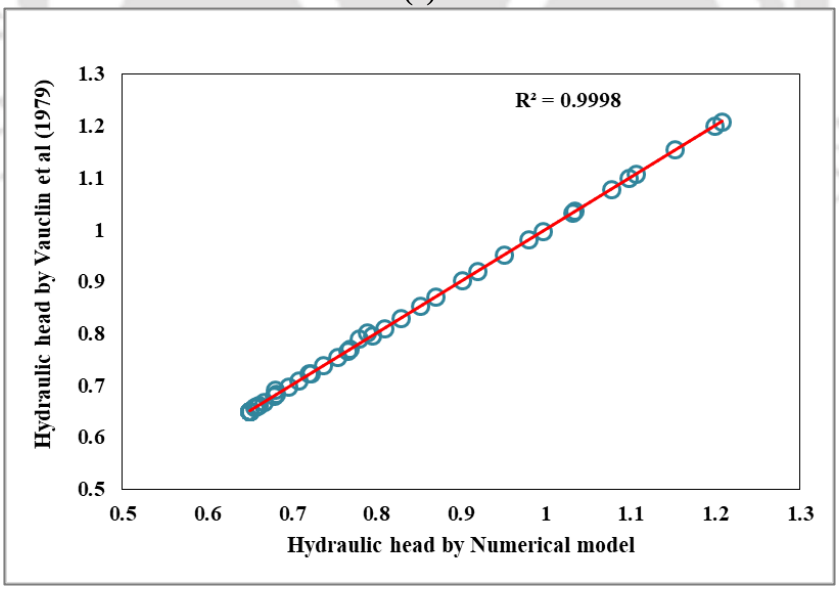
(b)

Figure 3. 5: Comparison of the pressure head distribution obtained from Paniconi et al. (1991) and Numerical simulation flow model.

The second model used for validation is a two-dimensional variably saturated zone in transient conditions. This is a problem of water table recharge on an area of 6m × 2m. Initially, the model is assumed to have the Groundwater table (GWT) located at a height of 0.65m, and a flux of 0.14791 m/ hr is applied at the top layer with an area of 1.00 m^2 at the center. Since the flow domain is symmetrical, only one-fourth of the model is simulated with the parameters given by Dogan & Motz, (2005). The comparison of the GWT position obtained from the simulation model and that obtained from the experimental result is presented in Figure 3.6.



(a)



(b)

Figure 3.6: Comparison of the Hydraulic head distribution obtained from Vauclin et al (1979) and Numerical simulation flow model.

From Figure 3.6 (a) it is observed the numerical flow simulation model could replicate an accurate solution as was obtained in the example problems of Vauclin et al (1979).The

coefficient of regression is found to be 0.9998, which ascertains that the numerical simulation flow model could provide an accurate solution as obtained by (Vauclin et al., 1979).The model validation using two example suggests that the model is a good representation of the physical system and can be used to make predictions or analyze scenarios where it may not be feasible or practical to conduct experiments or subsurface observations.

3.4 Methodology for Virus transport model:

This study develops the three-dimensional virus transport model in an unsaturated-saturated groundwater aquifer by considering variable attachment and detachment rates. The virus transport model is governed by four processes: Advection, Dispersion, Adsorption, and Inactivation. The governing equation discussed in section 3.4.1considers all the four processes that governs the virus transport in an aquifer.

3.4.1 Governing Equation:

The general equations governing virus transport in variably saturated media, including terms accounting for sorption onto Soil-water interfaces (SWI) and Air Water Interfaces (AWI), as stated by Torkzaban et al., (2006) is given as follows

$$\frac{\partial \theta C_W}{\partial t} = \frac{\partial}{\partial x}\left[\theta D \frac{\partial C_W}{\partial x}\right] - \frac{\partial q C_W}{\partial x} - \mu_W \theta C_W - \gamma_S - \gamma_a \tag{3.28}$$

$$\frac{\partial \rho_b C_S}{\partial t} = \gamma_S - \mu_S \rho_b C_S = \theta k_{att}^S C_W - k_{\det}^S \rho_b C_S - \mu_S \rho_b C_S \tag{3.29}$$

$$\frac{\partial \theta_a C_a}{\partial t} = \gamma_a - \mu_a \theta_a C_a = \theta k_{att}^a C_W - k_{\det}^a \theta_a C_a - \mu_a \theta_a C_a \tag{3.30}$$

Where, the concentration of free virus per unit volume in water denoted by C_w and C_S represents the virus concentration sorbed onto Soil-water-Interfaces (SWI) and C_a for Air-water-Interfaces (AWI); $\boldsymbol{\theta_a}$ (dimensionless) is the air content The dispersion coefficient (D) represents the rate of virus dispersion in the medium. The velocity flux (q) indicates the rate of fluid flow obtained using the flow model. The bulk density of soil (ρ_b) is the mass of soil per unit volume. The inactivation rate coefficients associated with water (μ_w), SWI (μ_s), and AWI (μ_a) represent the rates at which the virus becomes inactive. $\boldsymbol{\gamma_s}$ and $\boldsymbol{\gamma_a}$ are rates of adsorption to SWI and AWI, respectively The rates of attachment (k_{att}^s and k_{att}^a) and detachment (k_{det}^s and k_{det}^a) are associated with the SWI and AWI.

Zhang et al., (2012) developed an alternative and simpler formulation of the transport model by assuming attachment – detachment at the AWI occurs as a function of the available air-

water interfacial area. They obtained the governing equation by adding equation (3.28) and (3.30) to eliminate γ_a and the resulting equation becomes.

$$\frac{\partial \theta C_W}{\partial t} + \frac{\partial a C_{aW}}{\partial t} = \frac{\partial}{\partial x}\left[\theta D \frac{\partial C_W}{\partial x}\right] - \frac{\partial q C_W}{\partial x} - \mu_W \theta C_W - \gamma_S - \mu_a a C_{aW} \tag{3.31a}$$

$$\frac{\partial \theta C_W}{\partial t} + \frac{\partial a C_{aW}}{\partial t} = \frac{\partial}{\partial x}\left[\theta D \frac{\partial C_W}{\partial x}\right] - \frac{\partial q C_W}{\partial x} - \mu_W \theta C_W - \theta k_{att}^S C_W + k_{\det}^S \rho_b C_S - \mu_a a C_{aW} \tag{3.31b}$$

Where, **'a'** is the specific air–water interfacial area. In this study we considered relation between the interfacial area (a) and the degree of fluid saturation ($S_w = \theta/\theta_s$) by using the expression, $a = a_0 (1 - S_w)^3 S_w$, in which a_0 [$L^2 L^{-3}$] is the specific interfacial area corresponding to residual saturation of the medium.

A linear equilibrium partitioning of viruses between the water and air-water interface is assumed

$$C_{AW} = K_D^a C_W \tag{3.32}$$

Where, K_D^a is the equilibrium distribution coefficient for sorption onto the AWI expressed as:

$$K_D^a = \theta k_{att}^a / a k_{\det}^a \tag{3.33}$$

Using equation (3.32) and (3.33) we can rewrite the equation (3.31b) as

$$\frac{\partial \theta R C_W}{\partial t} = \frac{\partial}{\partial x}\left[\theta D \frac{\partial C_W}{\partial x}\right] - \frac{\partial q C_W}{\partial x} - \theta k_{att}^S C_W + k_{\det}^S \rho_b C_S - \mu_t \theta C_W + q_e C^* \tag{3.34}$$

$$R = 1 + \frac{a K_D^a}{\theta}$$

$$\mu_t = \mu_W + \frac{\mu_a a K_D^a}{\theta}$$

Where, the retardation coefficient (R) represents the extent of virus retardation in the medium. The additional qe and C^* is added which is nothing but the input source fluxes and source strength. The source strength of virus contamination (S_f) represents the amount of virus released into the system which is represented by C^* and the input source flux(If) is the fluxes injected in m/h which is denoted by qe .

By solving equation (3.34) with equation (3.29), the solution for one-dimensional virus transport in variably saturated zone using transient condition is obtained. This model is extended further into three-dimensional space as given by equation (3.35) and (3.36). The model developed using this equation will consider variable detachment rates based on the degree of saturation and thus, the model can be termed as 'variable detachment model'.

$$\frac{\partial \theta R C_W}{\partial t} = \frac{\partial}{\partial x_i}\left[\theta D_{ij}\frac{\partial C_W}{\partial x_j}\right] - \frac{\partial q_i C_W}{\partial x_i} - \theta k_{att}^S C_W + k_{\det}^S \rho_b C_S - \mu_t \theta C_W + q_e S_f \tag{3.35}$$

$$\frac{\partial \rho_b C_S}{\partial t} = \theta k_{att}^S C_W - k_{\det}^S \rho_b C_S - \mu_S \rho_b C_S \tag{3.36}$$

The virus transport model developed in this study is termed as 'Variable-detachment model'. The model uses varying adsorption rates based on the moisture velocity. In this study we have considered transport parameters for two saturation rate (65% and 100%) as available in the literature (Torkzaban et al. 2006a; 2006b), this model considers transport parameters of 65% saturation for unsaturated and 100 % for the saturation zones. Thus, for variable detachment rates, the following conditions is applied,

If $\theta < \theta_S$

$$k_{att}^S = k_{att}^{S,U}$$
$$k_{\det}^S = k_{\det}^{S,U} \quad ; \text{ U represents for unsaturated condition} \tag{3.37a}$$

If $\theta \geq \theta_S$

$$k_{att}^S = k_{att}^{S,S}$$
$$k_{\det}^S = k_{\det}^{S,S} \quad ; \text{ S represents for saturated condition} \tag{3.37b}$$
$$K_D^a = 0;$$

3.4.2 Numerical approximation:

The numerical scheme is considered in the similar way as it is used to develop the flow model in section 3.3.2, this model uses the implicit block-centered finite difference scheme. The discretization involved in solving the virus transport equation is discussed as below:

Advection term:

The advection term of the equation (3.35) can be expressed in three-dimensions and approximated through finite-difference as:

$$\frac{\partial (qC_W)_i}{\partial x_i} = \left[\frac{\partial}{\partial x}(q_x C_W) + \frac{\partial}{\partial y}(q_y C_W) + \frac{\partial}{\partial z}(q_z C_W)\right] \tag{3.38a}$$

$$= \left[\begin{array}{l} \frac{1}{\Delta x}\left(q_{x_{i+1/2,j,k}} C_{W_{i+1/2,j,k}}^{t+1} - q_{x_{i-1/2,j,k}} C_{W_{i-1/2,j,k}}^{t+1}\right) + \frac{1}{\Delta y}\left(q_{y_{i,j+1/2,k}} C_{W_{i,j+1/2,k}}^{t+1} - q_{y_{i,j-1/2,k}} C_{W_{i,j-1/2,k}}^{t+1}\right) \\ + \frac{1}{\Delta z}\left(q_{z_{i,j,k+1/2}} C_{W_{i,j,k+1/2}}^{t+1} - q_{z_{i,j,k-1/2}} C_{W_{i,j,k-1/2}}^{t+1}\right) \end{array}\right] \tag{3.38b}$$

Where, the concentration along the faces in x-axis is given by equation (3.39).

$$C^{t+1}_{W_{i+1/2,j,k}} = \delta_{i+1/2,j,k} C^{t+1}_{W_{i,j,k}} + \left(1-\delta_{i+1/2,j,k}\right) C^{t+1}_{W_{i+1,j,k}} \tag{3.39a}$$

$$C^{t+1}_{W_{i-1/2,j,k}} = \delta_{i-1/2,j,k} C^{t+1}_{W_{i-1,j,k}} + \left(1-\delta_{i-1/2,j,k}\right) C^{t+1}_{W_{i,j,k}} \tag{3.39b}$$

In this study, the upstream weighting scheme is used as a weighting factor. Such that $\delta_{i+1/2}$, $\delta_{j+1/2}$ and $\delta_{k+1/2}$ are the spatial weighting factor used in the advection term.

$$\delta_{i+1/2} = \begin{cases} 1 & if\ q_{i+1/2} > 0 \\ 0 & if\ q_{i+1/2} < 0 \end{cases} \tag{3.40a}$$

$$\delta_{i-1/2} = \begin{cases} 1 & if\ q_{i-1/2} > 0 \\ 0 & if\ q_{i-1/2} < 0 \end{cases} \tag{3.40b}$$

Therefore, the advection along y and z direction can be calculated in a similar way. The final advection term considering the components along x, y and z axis can be written as-

$$\frac{\partial\left(qC_W\right)_i}{\partial x_i} = \left[\begin{array}{l} q_{x_{i+1/2,j,k}} \left(\dfrac{\delta_{i+1/2,j,k} C^{t+1}_{W_{i,j,k}} + \left(1-\delta_{i+1/2,j,k}\right) C^{t+1}_{W_{i+1,j,k}}}{\Delta x_i} \right) - q_{x_{i-1/2,j,k}} \left(\dfrac{\delta_{i-1/2,j,k} C^{t+1}_{W_{i-1,j,k}} + \left(1-\delta_{i-1/2,j,k}\right) C^{t+1}_{W_{i,j,k}}}{\Delta x_i} \right) \\ + q_{y_{i,j+1/2,k}} \left(\dfrac{\delta_{i,j+1/2,k} C^{t+1}_{W_{i,j,k}} + \left(1-\delta_{i,j+1/2,k}\right) C^{t+1}_{W_{i,j+1,k}}}{\Delta y_j} \right) - q_{y_{i,j-1/2,k}} \left(\dfrac{\delta_{i,j-1/2,k} C^{t+1}_{W_{i,j-1,k}} + \left(1-\delta_{i,j-1/2,k}\right) C^{t+1}_{W_{i,j,k}}}{\Delta y_j} \right) \\ + q_{z_{i,j,k+1/2}} \left(\dfrac{\delta_{i,j,k+1/2} C^{t+1}_{W_{i,j,k}} + \left(1-\delta_{i,j,k+1/2}\right) C^{t+1}_{W_{i,j,k+1}}}{\Delta z_k} \right) - q_{z_{i,j,k-1/2}} \left(\dfrac{\delta_{i,j,k-1/2} C^{t+1}_{W_{i,j,k-1}} + \left(1-\delta_{i,j,k-1/2}\right) C^{t+1}_{W_{i,j,k}}}{\Delta z_k} \right) \end{array}\right] \tag{3.41}$$

Dispersion term:

The dispersion term of the equation (3.35) combines the molecular diffusion given by D^*_{ij} and the mechanical dispersion term called D^m_{ij}. This mechanical dispersion tensor is represented by a much more complex term called dispersivity, which is a 4thorder tensor in three dimensions. The dispersivity tensor (α_{ijkl}) and the Dispersion tensor (D) are related, as shown in equation (3.42), with the vector 'v' being the velocity vector. The dispersivity term can be simplified in practical scenarios to two components when the tensor is isotropic. These components are known as longitudinal dispersivity (α_L) and transverse dispersivity (α_T). However, the tensor is defined with three components, as shown in equation (3.43) (Burnett & Frind, 1987). The terms α_L, α_{TH}, and α_{TV} are longitudinal, transverse horizontal, and transverse

vertical dispersivity. *v*is the velocity vector considering the coordinate axes to be in line with the principal direction with the hydraulic conductivity tensor. The velocity vector (*v*) is nothing but $\left(q/\theta\right)$ and in the unsaturated zone, the movement of viruses due to molecular diffusion is generally slower compared to advective transport (caused by flowing groundwater). Therefore, in this study the molecular diffusion is neglected. The evaluation of the velocity component at cell interfaces in x-direction is shown in Figure 3.7.

$$D_{ij} = D_{ij}^{m} + D_{ij}^{*} = \alpha_{ijkl}\frac{v_{ik}v_{jl}}{|v|} + D_{ij}^{*} \tag{3.42}$$

$$D_{ij}^{m} = \begin{pmatrix} \alpha_L\frac{v_x^2}{|v|} + \alpha_{TH}\frac{v_y^2}{|v|} + \alpha_{TV}\frac{v_z^2}{|v|} & (\alpha_L - \alpha_{TH})\frac{v_x v_y}{|v|} & (\alpha_L - \alpha_{TV})\frac{v_x v_z}{|v|} \\ (\alpha_L - \alpha_{TH})\frac{v_x v_y}{|v|} & \alpha_{TH}\frac{v_x^2}{|v|} + \alpha_L\frac{v_y^2}{|v|} + \alpha_{TV}\frac{v_z^2}{|v|} & (\alpha_L - \alpha_{TV})\frac{v_y v_z}{|v|} \\ (\alpha_L - \alpha_{TV})\frac{v_x v_z}{|v|} & (\alpha_L - \alpha_{TV})\frac{v_y v_z}{|v|} & \alpha_{TV}\frac{v_x^2}{|v|} + \alpha_{TV}\frac{v_y^2}{|v|} + \alpha_L\frac{v_z^2}{|v|} \end{pmatrix} \tag{3.43}$$

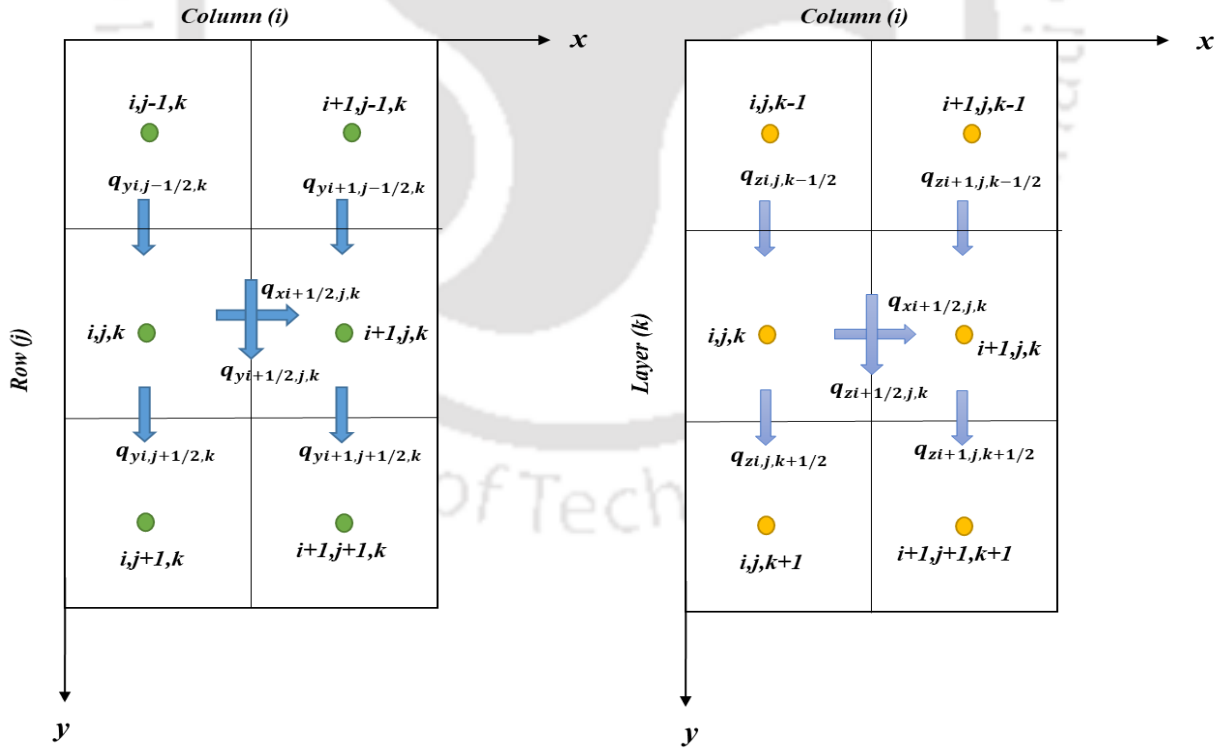


Figure 3.7: Evaluation of the velocity component at cell interfaces in x-direction for calculating the dispersion co-efficient D_{xx}, D_{yy} and D_{zz}.

$$\frac{\partial}{\partial x_i}\left[D_{ij}\frac{\partial C_W}{\partial x_j}\right]$$

$$=\frac{\partial}{\partial x}\left[\hat{D}_{xx}\frac{\partial C_W}{\partial x}\right]+\frac{\partial}{\partial x}\left[\hat{D}_{xy}\frac{\partial C_W}{\partial y}\right]+\frac{\partial}{\partial x}\left[\hat{D}_{xz}\frac{\partial C_W}{\partial z}\right]+\frac{\partial}{\partial y}\left[\hat{D}_{yx}\frac{\partial C_W}{\partial x}\right]+\frac{\partial}{\partial y}\left[\hat{D}_{yy}\frac{\partial C_W}{\partial y}\right] \quad (3.44)$$

$$+\frac{\partial}{\partial y}\left[\hat{D}_{yz}\frac{\partial C_W}{\partial z}\right]+\frac{\partial}{\partial z}\left[\hat{D}_{zx}\frac{\partial C_W}{\partial x}\right]+\frac{\partial}{\partial z}\left[\hat{D}_{zy}\frac{\partial C_W}{\partial y}\right]+\frac{\partial}{\partial z}\left[\hat{D}_{zz}\frac{\partial C_W}{\partial z}\right]$$

Using the finite difference approximation, the equation (3.44) can be expressed as given by equation (3.45)

$$\frac{\partial}{\partial x}\left[\hat{D}_{xx}\frac{\partial C_W}{\partial x}\right]=\hat{D}_{xx_{i+1/2,j,k}}\frac{C_{W\,i+1,j,k}^{t+1}-C_{W\,i,j,k}^{t+1}}{\Delta x_i\left(0.5\Delta x_i+0.5\Delta x_{i+1}\right)}-\hat{D}_{xx_{i-1/2,j,k}}\frac{C_{W\,i,j,k}^{t+1}-C_{W\,i-1,j,k}^{t+1}}{\Delta x_i\left(0.5\Delta x_i+0.5\Delta x_{i-1}\right)} \quad (3.45a)$$

$$\frac{\partial}{\partial y}\left[\hat{D}_{yy}\frac{\partial C_W}{\partial y}\right]=\hat{D}_{yy_{i,j+1/2,k}}\frac{C_{W\,i,j+1,k}^{t+1}-C_{W\,i,j,k}^{t+1}}{\Delta y_j\left(0.5\Delta y_j+0.5\Delta y_{j+1}\right)}-\hat{D}_{yy_{i,j-1/2,k}}\frac{C_{W\,i,j,k}^{t+1}-C_{W\,i,j-1,k}^{t+1}}{\Delta y_j\left(0.5\Delta y_j+0.5\Delta y_{j-1}\right)} \quad (3.45b)$$

$$\frac{\partial}{\partial z}\left[\hat{D}_{zz}\frac{\partial C_W}{\partial z}\right]=\hat{D}_{zz_{i,j,k+1/2}}\frac{C_{W\,i,j,k+1}^{t+1}-C_{W\,i,j,k}^{t+1}}{\Delta z_k\left(0.5\Delta z_k+0.5\Delta z_{k+1}\right)}-\hat{D}_{zz_{i,j,k-1/2}}\frac{C_{W\,i,j,k}^{t+1}-C_{W\,i,j,k-1}^{t+1}}{\Delta z_k\left(0.5\Delta z_k+0.5\Delta z_{k-1}\right)} \quad (3.45c)$$

$$\frac{\partial}{\partial x}\left[\hat{D}_{xy}\frac{\partial C_W}{\partial y}\right]=\hat{D}_{xy_{i+1/2,j,k}}\frac{\omega_{x_{i+1/2}}C_{W\,i,j+1,k}^{t+1}+\left(1-\omega_{x_{i+1/2}}\right)C_{W\,i+1,j+1,k}^{t+1}-\omega_{x_{i+1/2}}C_{W\,i,j-1,k}^{t+1}-\left(1-\omega_{x_{i+1/2}}\right)C_{W\,i+1,j-1,k}^{t+1}}{\Delta x_i\left(0.5\Delta y_{j-1}+\Delta y_j+0.5\Delta y_{j+1}\right)}$$

$$-\hat{D}_{xy_{i-1/2,j,k}}\frac{\omega_{x_{i-1/2}}C_{W\,i-1,j+1,k}^{t+1}+\left(1-\omega_{x_{i-1/2}}\right)C_{W\,i,j+1,k}^{t+1}-\omega_{x_{i-1/2}}C_{W\,i-1,j-1,k}^{t+1}-\left(1-\omega_{x_{i-1/2}}\right)C_{W\,i,j-1,k}^{t+1}}{\Delta x_i\left(0.5\Delta y_{j-1}+\Delta y_j+0.5\Delta y_{j+1}\right)} \quad (3.45d)$$

$$\frac{\partial}{\partial x}\left[\hat{D}_{xz}\frac{\partial C_W}{\partial z}\right]=\hat{D}_{xz_{i+1/2,j,k}}\frac{\omega_{x_{i+1/2}}C_{W\,i,j,k+1}^{t+1}+\left(1-\omega_{x_{i+1/2}}\right)C_{W\,i+1,j,k+1}^{t+1}-\omega_{x_{i+1/2}}C_{W\,i,j,k-1}^{t+1}-\left(1-\omega_{x_{i+1/2}}\right)C_{W\,i+1,j,k-1}^{t+1}}{\Delta x_i\left(0.5\Delta z_{k-1}+\Delta z_k+0.5\Delta z_{k+1}\right)}$$

$$-\hat{D}_{xz_{i-1/2,j,k}}\frac{\omega_{x_{i-1/2}}C_{W\,i-1,j,k+1}^{t+1}+\left(1-\omega_{x_{i-1/2}}\right)C_{W\,i,j,k+1}^{t+1}-\omega_{x_{i-1/2}}C_{W\,i-1,j,k-1}^{t+1}-\left(1-\omega_{x_{i-1/2}}\right)C_{W\,i,j,k-1}^{t+1}}{\Delta x_i\left(0.5\Delta z_{k-1}+\Delta z_k+0.5\Delta z_{k+1}\right)} \quad (3.45e)$$

$$\frac{\partial}{\partial y}\left[\hat{D}_{yx}\frac{\partial C_W}{\partial x}\right]=\hat{D}_{yx_{i,j+1/2,k}}\frac{\omega_{y_{j+1/2}}C_{W\,i+1,j,k}^{t+1}+\left(1-\omega_{y_{j+1/2}}\right)C_{W\,i+1,j+1,k}^{t+1}-\omega_{y_{j+1/2}}C_{W\,i-1,j,k}^{t+1}-\left(1-\omega_{y_{j+1/2}}\right)C_{W\,i-1,j+1,k}^{t+1}}{\Delta y_j\left(0.5\Delta x_{i-1}+\Delta x_i+0.5\Delta x_{i+1}\right)}$$

$$-\hat{D}_{yx_{i,j-1/2,k}}\frac{\omega_{x_{j-1/2}}C_{W\,i+1,j-1,k}^{t+1}+\left(1-\omega_{x_{j-1/2}}\right)C_{W\,i+1,j,k}^{t+1}-\omega_{x_{j-1/2}}C_{W\,i-1,j-1,k}^{t+1}-\left(1-\omega_{x_{j-1/2}}\right)C_{W\,i-1,j,k}^{t+1}}{\Delta y_j\left(0.5\Delta x_{i-1}+\Delta x_i+0.5\Delta x_{i+1}\right)} \quad (3.45f)$$

$$\frac{\partial}{\partial y}\left[\hat{D}_{yz}\frac{\partial C_W}{\partial z}\right]=\hat{D}_{yz_{i,j+1/2,k}}\frac{\omega_{y_{j+1/2}}C_{W\,i,j,k+1}^{t+1}+\left(1-\omega_{y_{j+1/2}}\right)C_{W\,i,j+1,k+1}^{t+1}-\omega_{y_{j+1/2}}C_{W\,i,j,k-1}^{t+1}-\left(1-\omega_{y_{j+1/2}}\right)C_{W\,i,j+1,k-1}^{t+1}}{\Delta y_j\left(0.5\Delta z_{k-1}+\Delta z_k+0.5\Delta z_{k+1}\right)}$$

$$-\hat{D}_{yz_{i,j-1/2,k}}\frac{\omega_{x_{j-1/2}}C_{W\,i,j-1,k+1}^{t+1}+\left(1-\omega_{x_{j-1/2}}\right)C_{W\,i,j,k+1}^{t+1}-\omega_{x_{j-1/2}}C_{W\,i,j-1,k-1}^{t+1}-\left(1-\omega_{x_{j-1/2}}\right)C_{W\,i,j,k-1}^{t+1}}{\Delta y_j\left(0.5\Delta z_{k-1}+\Delta z_k+0.5\Delta z_{k+1}\right)} \tag{3.45g}$$

$$\frac{\partial}{\partial z}\left[\hat{D}_{zx}\frac{\partial C_W}{\partial x}\right]=\hat{D}_{zx_{i,j,k+1/2}}\frac{\omega_{z_{k+1/2}}C_{W\,i+1,j,k}^{t+1}+\left(1-\omega_{z_{k+1/2}}\right)C_{W\,i+1,j,k+1}^{t+1}-\omega_{z_{k+1/2}}C_{W\,i-1,j,k}^{t+1}-\left(1-\omega_{z_{k+1/2}}\right)C_{W\,i-1,j,k+1}^{t+1}}{\Delta z_k\left(0.5\Delta x_{i-1}+\Delta x_i+0.5\Delta x_{i+1}\right)}$$

$$-\hat{D}_{zx_{i,j,k-1/2}}\frac{\omega_{z_{k-1/2}}C_{W\,i+1,j,k-1}^{t+1}+\left(1-\omega_{z_{k-1/2}}\right)C_{W\,i+1,j,k}^{t+1}-\omega_{z_{k-1/2}}C_{W\,i-1,j,k-1}^{t+1}-\left(1-\omega_{z_{k-1/2}}\right)C_{W\,i-1,j,k}^{t+1}}{\Delta z_k\left(0.5\Delta x_{i-1}+\Delta x_i+0.5\Delta x_{i+1}\right)} \tag{3.45h}$$

$$\frac{\partial}{\partial z}\left[\hat{D}_{zy}\frac{\partial C_W}{\partial y}\right]=\hat{D}_{zy_{i,j,k+1/2}}\frac{\omega_{z_{k+1/2}}C_{W\,i,j+1,k}^{t+1}+\left(1-\omega_{z_{k+1/2}}\right)C_{W\,i,j+1,k+1}^{t+1}-\omega_{z_{k+1/2}}C_{W\,i,j-1,k}^{t+1}-\left(1-\omega_{z_{k+1/2}}\right)C_{W\,i,j-1,k+1}^{t+1}}{\Delta z_k\left(0.5\Delta y_{j-1}+\Delta y_j+0.5\Delta y_{j+1}\right)}$$

$$-\hat{D}_{zy_{i,j,k-1/2}}\frac{\omega_{z_{k-1/2}}C_{W\,i,j+1,k-1}^{t+1}+\left(1-\omega_{z_{k-1/2}}\right)C_{W\,i,j+1,k}^{t+1}-\omega_{z_{k-1/2}}C_{W\,i,j-1,k-1}^{t+1}-\left(1-\omega_{z_{k-1/2}}\right)C_{W\,i,j-1,k}^{t+1}}{\Delta z_k\left(0.5\Delta y_{j-1}+\Delta y_j+0.5\Delta y_{j+1}\right)} \tag{3.45i}$$

Where, ω_x, ω_y and ω_z are the spatial weighting factors used to compute the concentration value at a cell interface when evaluating a dispersion cross term, which is defined as:

$$\omega_{x_{i+1/2}}=\frac{\Delta x_{i+1}}{\Delta x_i+\Delta x_{i+1}};\omega_{y_{j+1/2}}=\frac{\Delta y_{j+1}}{\Delta y_j+\Delta y_{j+1}};\omega_{z_{k+1/2}}=\frac{\Delta z_{k+1}}{\Delta z_k+\Delta z_{k+1}}$$

Sorption Terms:

In the virus transport equation, the adsorption coefficient is typically represented as a function of time and distance and the concentration of viruses in the groundwater. This allows for the simulation of the time-dependent behavior of the adsorption process and its effects on the transport of viruses in the aquifer. Attachment and detachment are the two processes that governs the kinetic sorption in the virus transport model. Kinetic sorption is a commonly used model for virus transport studies, as it takes into account the non-equilibrium conditions that often occur in real-world scenarios. In this model, viruses' attachment and detachment rates are considered rather than assuming an equilibrium condition. The virus sorption is given by equation (3.36), which is expanded into finite difference form represented by equation (3.46).

$$\rho_b\left[\frac{C_{S_{ijk}}^{t+1}-C_{S_{ijk}}^{t}}{\Delta t}\right]=\theta_{ijk}^{t+1}k_{att}^{S}C_{W_{ijk}}^{t+1}-k_{\det}^{S}\rho_b C_{S_{ijk}}^{t+1}-\mu_S\rho_b C_{S_{ijk}}^{t+1} \tag{3.46a}$$

$$C_{S_{ijk}}^{t+1}=\frac{C_{S_{ijk}}^{t}+\Delta t/\rho_b\left[\theta_{ijk}^{t+1}k_{att}^{S}C_{W_{ijk}}^{t+1}\right]}{\left[1+\Delta t\left(k_{\det}^{S}+\mu_s\right)\right]} \tag{3.46b}$$

Now coupling equation (3.46b) in the equation (3.35) we get the following expression

$$\frac{\partial \theta R C_W}{\partial t} = L(C) - \theta_{ijk}^{t+1} k_{att}^{S} C_{W\,ijk}^{\;t+1} + k_{\det}^{S} \rho_b \left(\frac{C_{S_{ijk}}^{t} + \Delta t / \rho_b \left[\theta_{ijk}^{t+1} k_{att}^{S} C_{W_{ijk}}^{t+1} \right]}{\left[1 + \Delta t \left(k_{\det}^{S} + \mu_s \right) \right]} \right) - \mu_t \theta_{ijk}^{t+1} C_{W\,ijk}^{\;t+1} + q_e C^{*} \tag{3.47}$$

$$\frac{\partial \theta R C_W}{\partial t} = 3\theta_{ijk}^{t+1} C_{W\,ijk}^{\;t+1} R_{ijk}^{t+1} - \theta_{ijk}^{t+1} C_{W\,ijk}^{\;t+1} R_{ijk}^{t} - \theta_{ijk}^{t} C_{W\,ijk}^{\;t+1} R_{ijk}^{t+1} - \theta_{ijk}^{t+1} C_{W\,ijk}^{\;t} R_{ijk}^{t+1} \tag{3.48}$$

Here, L(C) represents the advection and dispersion term obtained from equation (3.41) and (3.45) respectively. The source term mainly represents the injection of virus into the aquifer and the sink term represents the rate of the fluid pumping out from the aquifer. C* represents the input virus source strength (*Sf*) and *qe* represents the input source flux (*If*) that enters into the groundwater aquifer. Substituting, all the components in the transport equation, and multiplying both sides by ($\Delta x \Delta y \Delta z$) we obtain the following finite difference form of a cell (*ijk*) given by equation (3.49).

$$\begin{aligned} &A_{ijk}^{1} C_{W\,ijk}^{\;t+1} + A_{ijk}^{2} C_{W\,i,j,k-1}^{\;t+1} + A_{ijk}^{3} C_{W\,i,j,k+1}^{\;t+1} + A_{ijk}^{4} C_{W\,i,j-1,k}^{\;t+1} + A_{ijk}^{5} C_{W\,i,j+1,k}^{\;t+1} + A_{ijk}^{6} C_{W\,i-1,j,k}^{\;t+1} + A_{ijk}^{7} C_{W\,i+1,j,k}^{\;t+1} \\ &+ A_{ijk}^{8} C_{W\,i,j-1,k-1}^{\;t+1} + A_{ijk}^{9} C_{W\,i,j+1,k-1}^{\;t+1} + A_{ijk}^{10} C_{W\,i,j-1,k+1}^{\;t+1} + A_{ijk}^{11} C_{W\,i,j+1,k+1}^{\;t+1} + A_{ijk}^{12} C_{W\,i-1,j,k-1}^{\;t+1} + A_{ijk}^{13} C_{W\,i+1,j,k-1}^{\;t+1} \\ &+ A_{ijk}^{14} C_{W\,i-1,j,k+1}^{\;t+1} + A_{ijk}^{15} C_{W\,i+1,j,k+1}^{\;t+1} + A_{ijk}^{16} C_{W\,i-1,j-1,k}^{\;t+1} + A_{ijk}^{17} C_{W\,i+1,j-1,k}^{\;t+1} + A_{ijk}^{18} C_{W\,i-1,j+1,k}^{\;t+1} + A_{ijk}^{19} C_{W\,i+1,j+1,k}^{\;t+1} = b_{i,j,k} \end{aligned} \tag{3.49}$$

Equation (3.48) can be expressed in matrix form as given by equation (3.49)

$$\left[A\right]^{t} C_{W}^{\;t+1} = \left\{RHS\right\} \tag{3.50}$$

Where [A] is a square matrix consisting of the coefficients (A^1 to A^{19}) of the finite difference equation (3.50), $C^{t+1}$is unknown virus concentration values for the current time step, and {RHS} is the forcing vector consisting of known values from the previous time step. This system of linear equations is solved using the preconditioned conjugate gradient method (PCGM). The fluid flow equation of the previous section also employed the PCGM method. However, the flow model used 7 cells to formulate the system of equations, whereas, the transport model employed 19 cells to formulate the matrix equation.

3.4.3 Initial and boundary condition for transport model:

Initial condition:

As an initial condition no concentration is present in the study domain. The general form of the initial condition is given by:

$$C_w(x, y, z, t) = C_0(x, y, z, t) \;; \qquad t \geq 0 \tag{3.51}$$

Where, C_0 is the known concentration and it is equal to zero in this study.

Boundary condition

For the Dirichlet boundary condition, the concentration is specified along the boundary for the entire duration of the simulation,

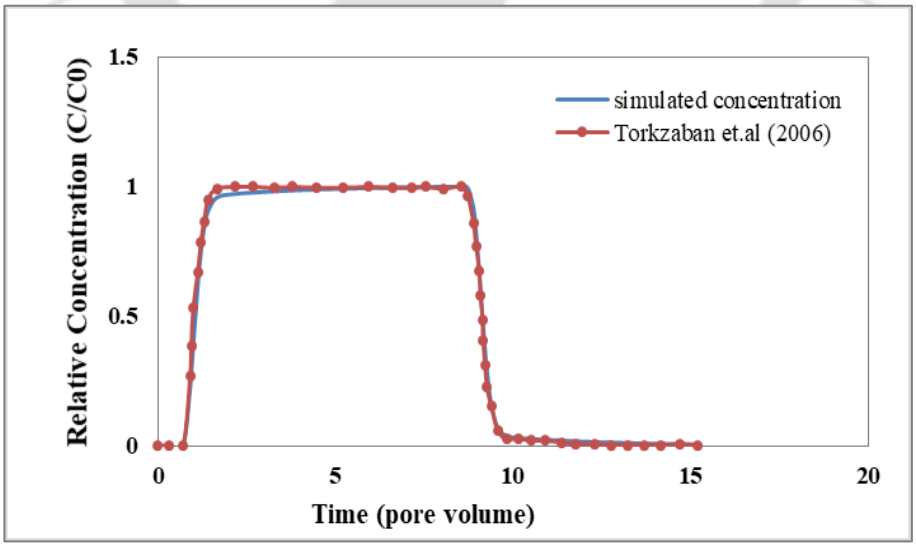
$$C_w(x,y,z,t) = c_w(x,y,z,t) \; ; \qquad t \geq 0 \tag{3.52}$$

For the impervious boundary, both dispersive and advective fluxes are equal to zero.

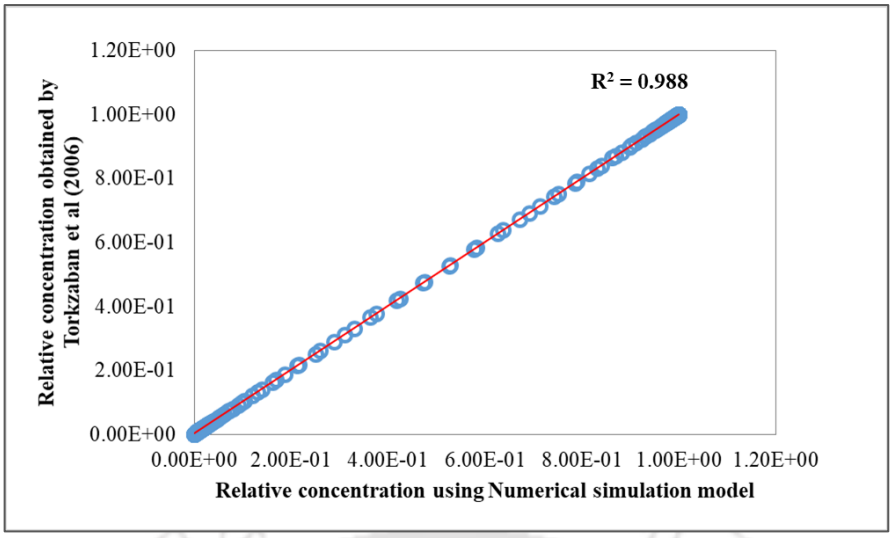
$$\frac{\partial C_w}{\partial x_j} = 0; \qquad t \geq 0 \tag{3.53}$$

3.4.4 Model validation:

Validation of the solutions obtained by the numerical virus transport model is a critical step to ensure its accuracy and reliability. One of the ways to validate the model is by comparing its results with available experimental data. Suppose the solutions obtained by the model are in good agreement with the experimental results. In that case, it can be concluded that the model is valid and can be used to make predictions for similar scenarios. Research here utilizes a one-dimensional column experiment under unsaturated flow conditions conducted by Torkzaban et al. (2006) for validation. The column length is 23 cm filled with a homogeneous soil in it and the microbe considered in this study is the MS2 virus. A constant saturation of 65% is maintained throughout the length of the model, the same as it was set up for the column experiment of Torkzaben et al. (2006). The virus source was made inactive as soon as the fluid flow in the column reached a steady state. The boundary conditions and parameters used to develop the simulation model are the same as Torkzaban et al. (2006). The numerical model was employed for one dimensional spatial analysis for 40000 seconds and the breakthrough curve obtained from the analysis is plotted along with the experimental data (Figure 3. 8).



(a)



(b)

Figure 3. 8: Comparison of the breakthrough curve obtained from Torkzaban et al (2006) and Simulation virus transport model.

To validate this model, the three-dimensional virus transport model written in MATLAB is converted into one dimension by considering only the z-axis and the remaining x- and y- axis was kept equal to 1. The comparative plot (Figure 3. 8)indicates that the numerical model exactly fits the experimental data. The regression value is calculated and it was found that the R^2 value is equal to 0.988. This value is very close to 1, which indicates that the model developed for simulating the virus transport model can accurately simulate the virus transport in field scale.

3.5 Results and Discussion:

In this study two scenarios are generated to analyze the movement of virus from the unsaturated zone to the saturated zone. Initially, the flow model is developed and the flow simulation include the hydraulic head distribution, velocity field, and water table fluctuation over time. The solution provides the flow patterns and direction of groundwater movement in the aquifer, identifying areas of recharge, discharge, and potential flow pathways that can influence the virus transport. Once, the flow model is developed, the solutions of flow velocity and moisture content are incorporated into the virus transport model. The solution obtained from the virus transport simulation focuses on the distribution and behavior of the virus concentration over time. It shows the contour plots illustrating the spatial and temporal variation of virus concentration in the aquifer. A sensitivity analysis for the spatial and time discretization is also carried out and the best suitable time step and space size is chosen for developing the model

(ref. Figure 3. 9) Table 3. 1 and Table 3. 2 provides the input flow and transport parameters used to develop the virus transport model in unsaturated- saturated groundwater aquifer.

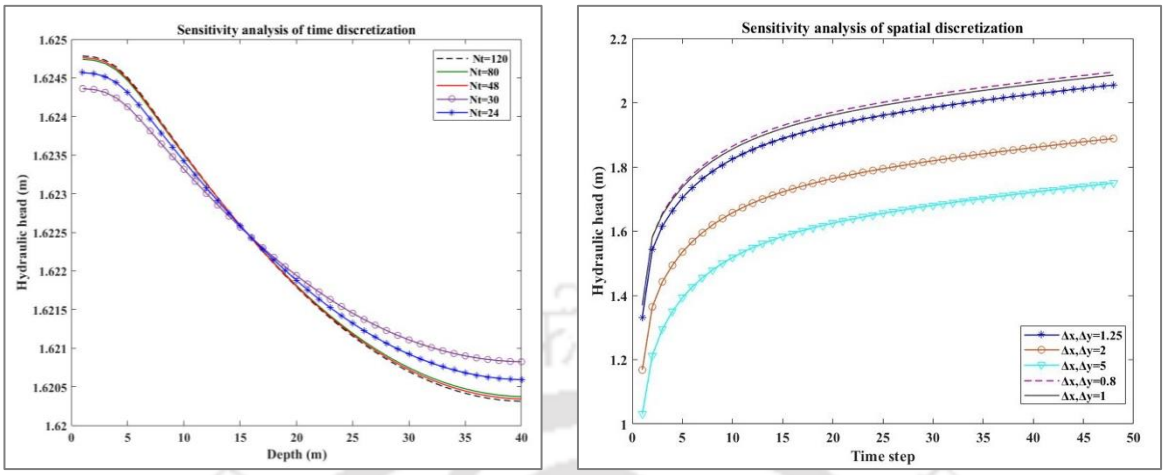


Figure 3. 9:Sensitivity analysis of time and spatial discretization.

Table 3. 1: Input parameters used to develop the groundwater flow model.

Sl. No	Parameters	Value
1	Flow domain	$20\times20\times2\,(m^3)$
2	Saturated Hydraulic conductivity (K_s)	0.678 mh^{-1}
3	Residual moisture content	0.055
4	Saturated moisture content	0.41
5	Air entry pressure ($h_a=1/\alpha$)	1/3.55m
6	Specific storage (S_s)	0.001 (m^{-1}) (assumed)
7	Van Genuchten parameter(n)	4.1
8	Bottom, front, back and left boundary of the flow domain	Impervious – no flow condition
9	Top boundary	Prescribed flux at top
10	Right boundary	The water level is maintained at 1.625m as a constant head boundary
11	Initial pressure head	Hydrostatic equilibrium with horizontal water table at 1.625m (i.e., h+z= 1.625m)
12	Grid discretization	cells with size of Δx=1 m, Δy = 1m, Δz = 0.05 m
13	Time increment	0.5time step is maintained
14	Maximum simulation time	24 hours

Table 3. 2: Input transport parameters of the virus transport model (Zhang et al., 2012)

	Parameters	units	Value
1	Type of virus	MS2 virus	
2	Longitudinal dispersivity (α_L)	m	4.4
3	Horizontal dispersivity (α_{TH})	m	0.44
4	Transverse dispersivity (α_{TV})	m	0.044
5	Saturated hydraulic conductivity	m h^{-1}	0.678
6	Soil dry bulk density (ρ_b)	g m^{-3}	1.65×10^{6}
7	Solid–water interface (SWI) attachment coefficient ($k_{att}^{s,u}$) in 65% saturation	h^{-1}	0.0696
8	SWI detachment coefficient ($k_{det}^{s,u}$) in 65% saturation	h^{-1}	0.504
9	Solid–water interface (SWI) attachment coefficient ($k_{att}^{s,s}$) in 100% saturation	h^{-1}	0.594
10	SWI detachment coefficient ($k_{det}^{s,s}$) in 100% saturation	h^{-1}	0.095
11	Inactivation rate (μ_w, μ_s, μ_a)	h^{-1}	7.79×10^{-4}
12	K_D[a]	m	0.04
13	a_0	m^{-1}	4×10^{5}

3.5.1 Numerical solution of Three-dimensional groundwater flow and transport model for Scenario 1:

The study utilizes the synthetic groundwater flow numerical model in three-dimensional space simulating unconfined aquifers having unsaturated and saturated zones. The model is solved for transient state conditions for homogeneous soil domains. The model has a dimension of (20m×20m×2m), and it was assumed that the water table is initially horizontal and 1.625m above the bottom of the aquifer. Two pumping wells are located, and the water is pumped out from the saturated zone. Two sources of virus contamination are taking place due to leakage from sewer lines in the unsaturated zone. The graphical representation of the study area is shown in Figure 3. 10. According to the description, the proposed flow model is highly nonlinear, as such a hydraulic process creates large gradients in soil pressure and soil moisture content during the process of infiltration. Therefore, selecting the appropriate grid size when developing a numerical model is very significant. Such a problem can be solved by choosing a variable grid size. When saturation is uneven, a small node increment across the z-axis can be used to save computational time and accurately represent the change in physical phenomena along the depth. In this scenario, the datum is located at the bottom of the flow domain, with Z=0 representing the bottom and Z=2m representing the top. A block-centered finite difference

scheme is being used, which means the nodes are positioned at the center of each grid cell. For a grid size of (1m×1m×0.05m), the nodes are situated at various depths relative to the ground surface. Specifically, the top layer is at a depth Z=1.975m, the second layer is at Z= 1.925m, the third layer is at Z=1.875m, and so on. The details of the grid size, hydraulic and soil properties used to develop the flow model are listed in Table 3. 1and Table 3. 2. Also, the detailed of the pumping well location and the source locations along with the magnitude of the source fluxes are listed in Table 3. 3.

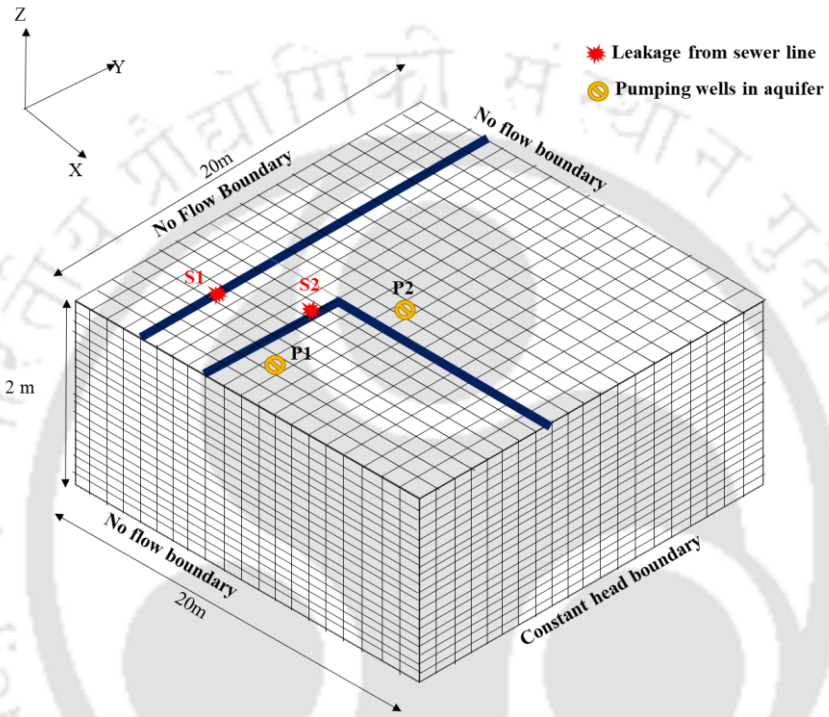


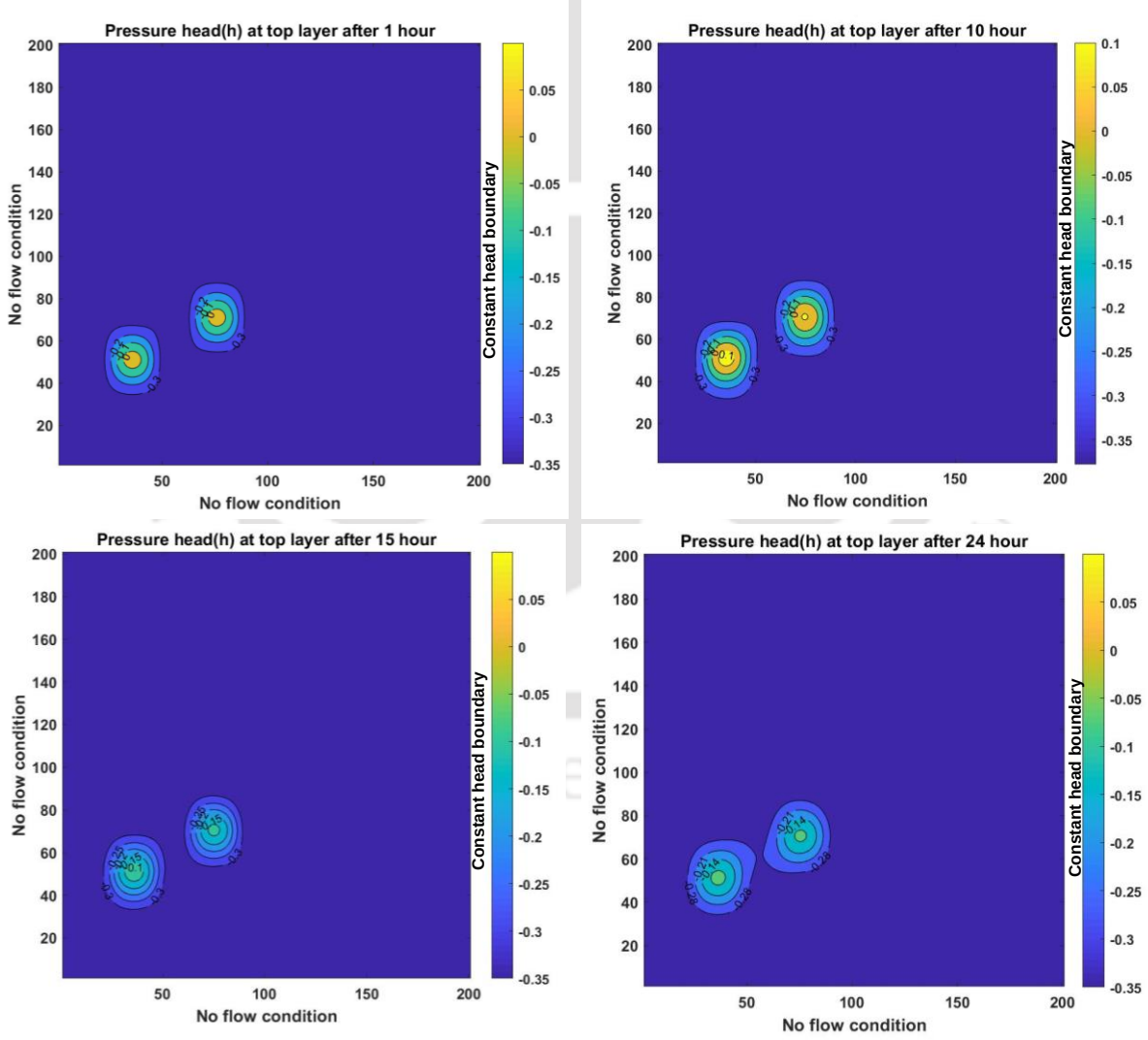
Figure 3. 10: Hypothetical study area for Scenario 1 showing the two sources location from the leakage from sewer lines.

Table 3. 3: Pumping well and source locations along with fluxes for scenario 1

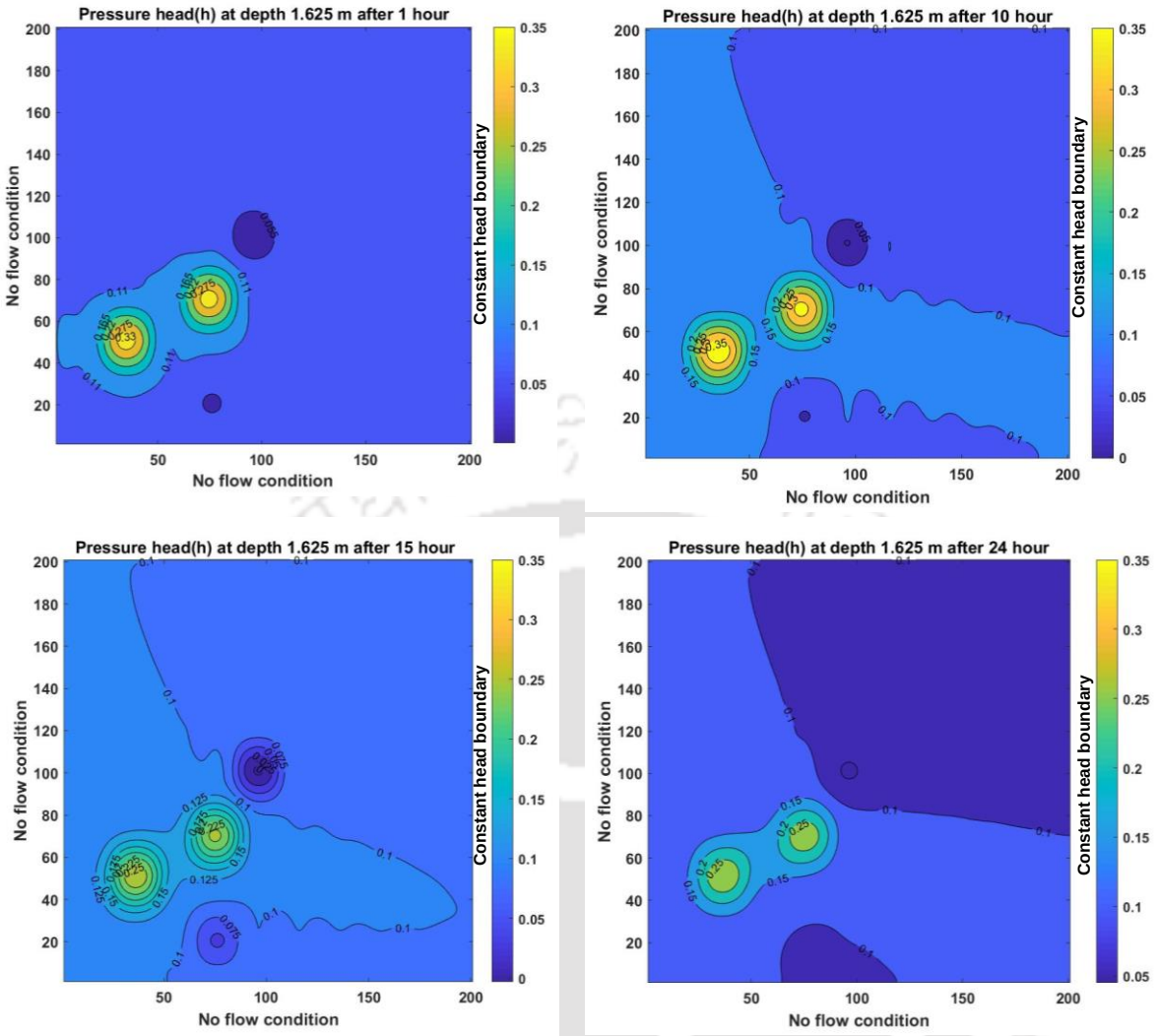
	Input data	Value	Active	inactive
1	Input source flux(If_1)	0.35 m/h	0-10 hours	10-24 hours
2	Constant flux(If_2)	0.37 m/h	0-10 hours	10-24 hours
3	Pumping rate(q_e)	-0.25 m/h	0-15 hours	15-24 hours
4	Source location 1 (Sl_1) in x,y,z c-ordinate axis	(4,5,1)	-	-
5	Source location 2 (Sl_2) in x,y,z c-ordinate axis	(8,7,1)	-	-
6	Source release strengths for location 1 (Sf_1)	400 g/h	0-10 hours	10-24 hours
7	Source release strengths for location 2 (Sf_2)	500 g/h	0-10 hours	10-24 hours

3.5.1.1 Variation of pressure head (m) obtained from the model 1:

Pressure head contour plots obtained from the groundwater flow model is presented in Figure 3. 11. This provides a visual representation of the distribution of pressure head values throughout the aquifer. These contour plots are usually displayed as a series of lines that connect points of equal pressure head values. In Figure 3. 11 (a), negative pressure head is seen which indicates the aquifer is in an unsaturated or partially saturated state. This condition is often associated with the vadose zone, which is the area above the water table. Within this area, soil moisture content is low indicating drier conditions in the unsaturated zone. In the source location, where there is recharge, positive pressure head is found. This indicates that, the pore spaces within the region are filled with water, and the pressure exerted by the water is greater than atmospheric pressure.



(a) Pressure head (m) in unsaturated zone



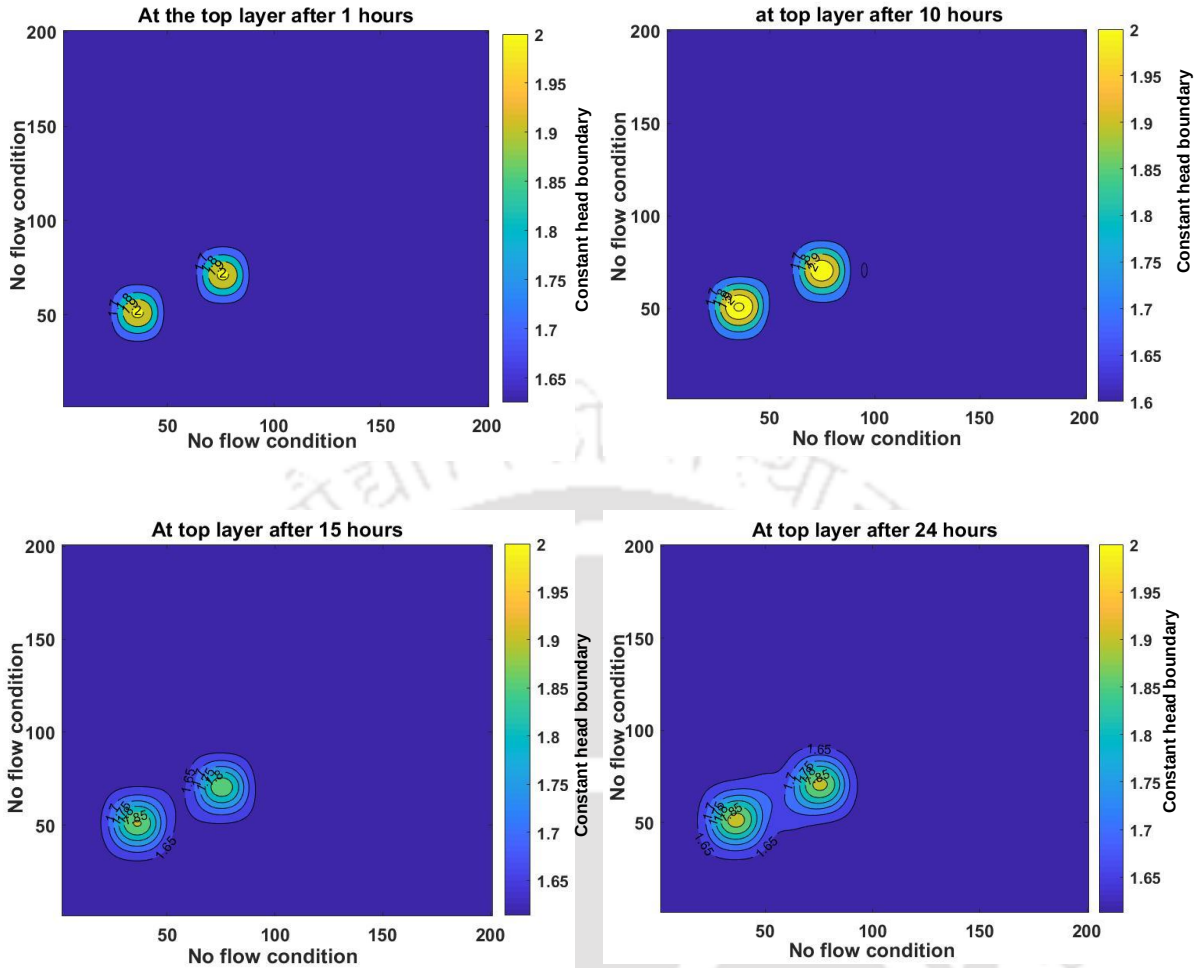
(b) Pressure head (m) in saturated zone

Figure 3. 11 : Contour plots of the pressure head distribution in (a) Unsaturated zone and (b) saturated zone.

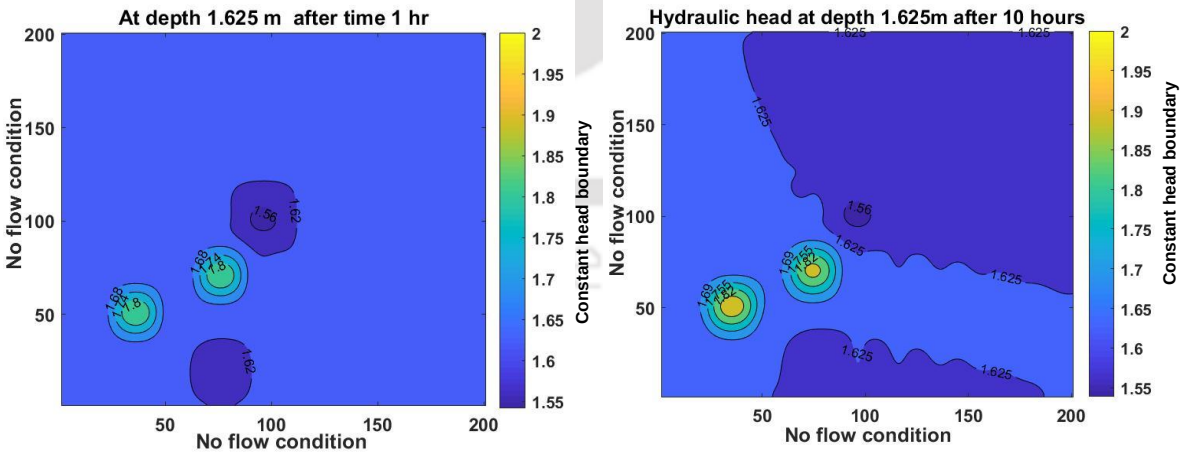
While referring the pressure head distribution in the saturated zone, the pumping well location creates a cone-shaped depression in the water table around the well. This cone of depression results in decrease in the pressure head in the immediate vicinity of the well as shown in Figure 3. 11 (b).

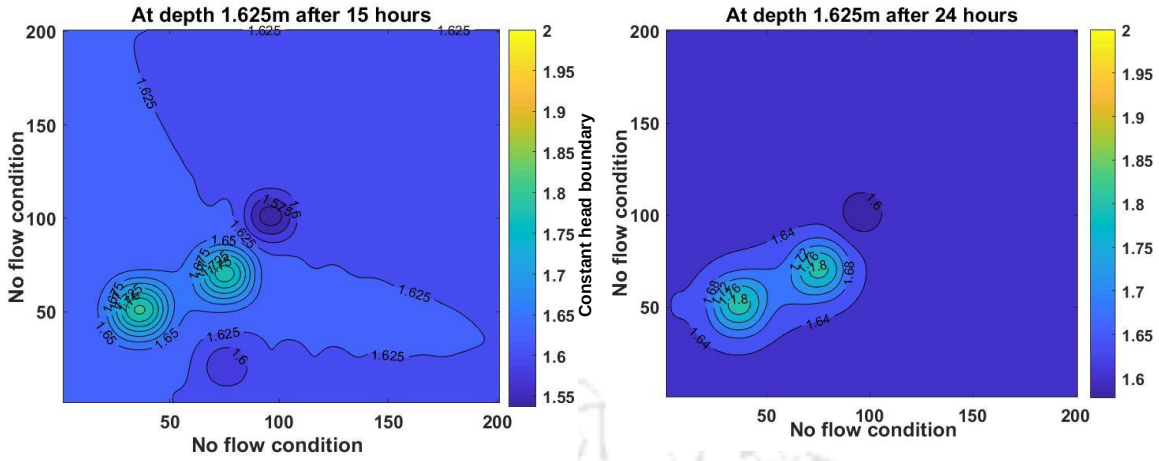
3.5.1.2 Variation of the hydraulic head (m) obtained from model 1:

The hydraulic head represents the combined effects of pressure head and elevation head. The hydraulic head in an unconfined aquifer can vary spatially and temporally due to factors such as recharge rates, groundwater pumping, and natural variations in precipitation. The variation in hydraulic head obtained from the groundwater simulation model developed for Scenario 1 is presented in Figure 3.12.



(a) Hydraulic head (m) in unsaturated zone





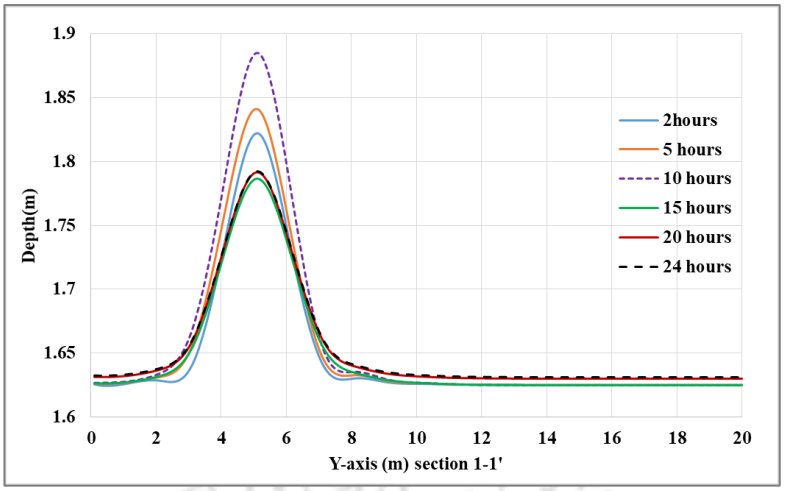
(b) Hydraulic head (m) in saturated zone

Figure 3.12: Contour plots showing the hydraulic head at different time steps in saturated and unsaturated zone.

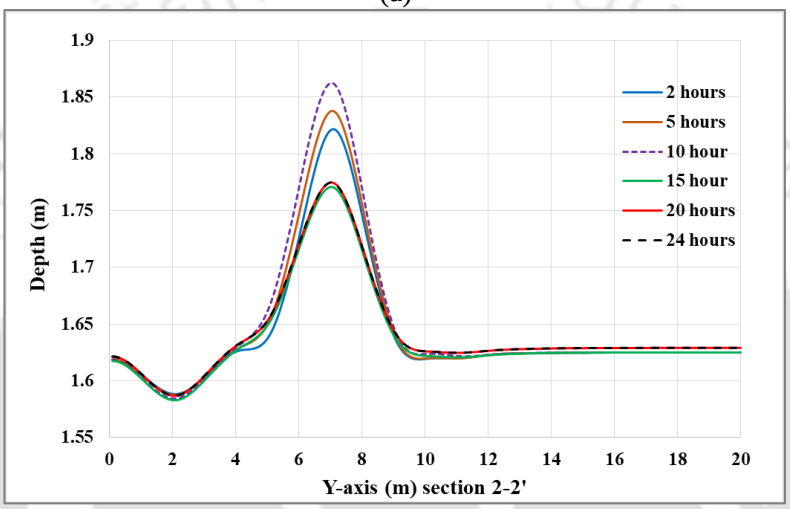
Referring to Figure 3.12, a cone of depression or drawdown in the water table is observed in the pumping well location. Initially, the water table was located at a depth of 1.625m from the datum, but after 15 hours of pumping, the hydraulic head has lowered to 1.57m in pumping well locations. This is because the pumping reduces pressure and draws water into the well. On the other hand, the input flux recharge in the source location adds moisture to the aquifer, increasing the water table elevation and raising the hydraulic head. It is found that the hydraulic head has raised to 1.875 m from the datum in the source locations. The details of the water table fluctuation with time are given in section 3.5.1.3.

3.5.1.3 Variation of water table position obtained from the model 1:

The variation in the water table due to pumping and recharge (described in Scenario-1) in the unconfined aquifer is modeled in this section of the study. As we know, the water table of an unconfined aquifer is subject to atmospheric pressure. The water table elevation can fluctuate both vertically and horizontally in response to various factors, including precipitation, evaporation, groundwater recharge, and groundwater pumping. The water table position plot is an important output of a groundwater flow model, as it provides valuable information about the distribution of groundwater in the aquifer. It can be used to identify areas of high and low water availability, as well as potential recharge and discharge zones. Figure 3.13 shows the water table position by proving a series of lines, with each line representing a different elevation of the water table at different time steps. The lines are labeled with the corresponding elevation in meters.



(a)



(b)

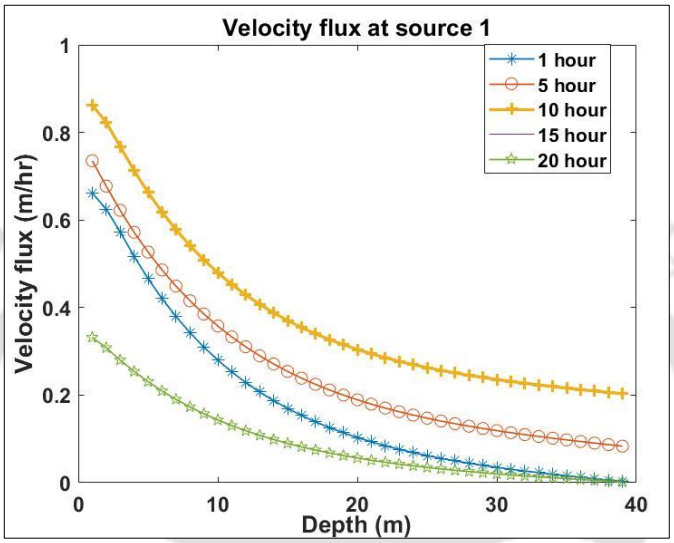
Figure 3.13: Water table position at different time steps at(a) section 1-1’ and (b) 2-2’ for scenario 1

In Figure 3.13 (a), the peaks in the graph represent a rise in the water table, indicating an increase in the amount of water stored in the aquifer. On the other hand, in Figure 3.13 (b) rise in water table is seen as well as the graph shows a falling trend which indicates a depletion of the groundwater table. Both in section 1-1’ and 2-2’, the rise in water table is observed in the source location at 10 hours of simulations, after that the water table has decreased at 15 hours. A cone-shaped depression in the water table around the pumping well 2 is seen in section 2-2’ indicating that the water table has lowered upto 1.575m. After 15 hours of simulation the water table fluctuation does not change and it remains constant.

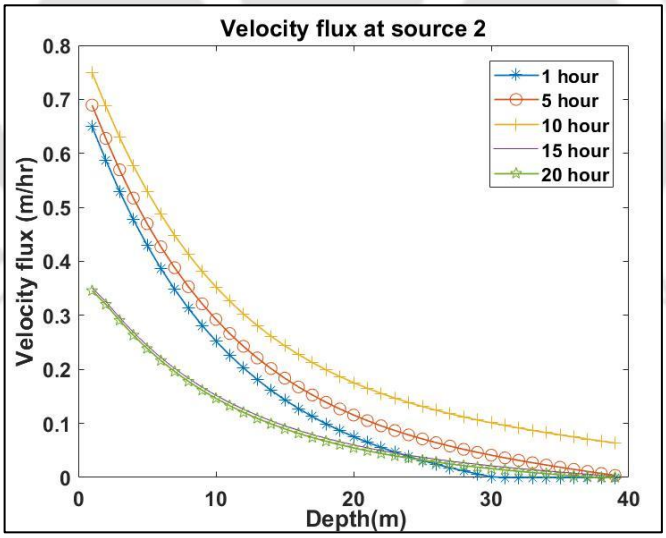
3.5.1.4 Velocity flux obtained from the model 1:

The spatial variation of velocity flux values across the two source locations of Scenario-1 is obtained after 24 hours of simulation and is shown in Figure 3. 14. The velocity flux increases in both sources during the first 10 hours of time period. This suggests an initial rise in the

groundwater flow rate due to the groundwater recharge in both locations. After the initial increase, the velocity flux starts to decrease until it reaches a lower value at 15 hours. This indicates a gradual reduction in groundwater flow rate in both sources during this time interval. This is due to the pumping effect, and the pumping well was kept active till 15 hours. Beyond 15 hours, the velocity flux remains constant. This implies that the rate of groundwater flow from both sources stabilizes, and there is no further change in the velocity flux values.



(a)



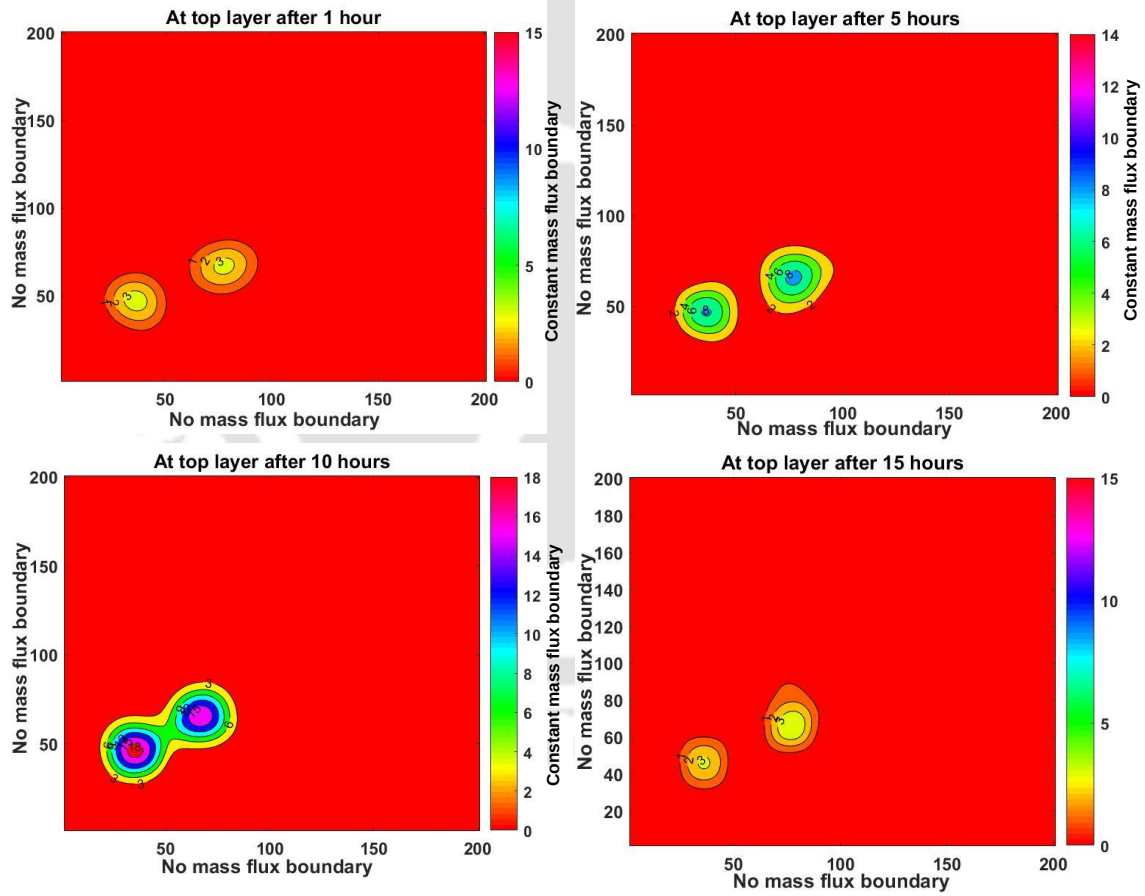
(b)

Figure 3. 14: Variation of velocity flux at two source location with depth at different time step in scenario 1.

3.5.1.5 Variation of virus concentration obtained from the model 1:

Contour plots of virus concentration in groundwater are graphical representations that provide a visual depiction of the distribution of virus concentration levels across a specific area or

region of groundwater. While generating the virus concentration at different locations and times, it is observed that the concentration data points are not evenly distributed across the entire study area. To generate a continuous contour plot, interpolation techniques is employed, which estimate the virus concentration values at unsampled locations based on the available data. With the dataset prepared and interpolated, the contour plot is generated. The concentration data obtained from the virus transport model for scenario 1 is presented in Figure 3. 15 and Figure 3. 16.The x and y axes of the plot represent the spatial coordinates, while the contour lines represent different virus concentration levels in g/h. This can be converted to a number of viruses per liter by multiplying with 10^6, which assumes the number of viruses for pathogenic wastewater as 1 billion viruses per gram of waste. Each contour line represents a constant value of virus concentration. Typically, lines that are closer together indicate higher concentrations, while lines that are further apart indicate lower concentrations.



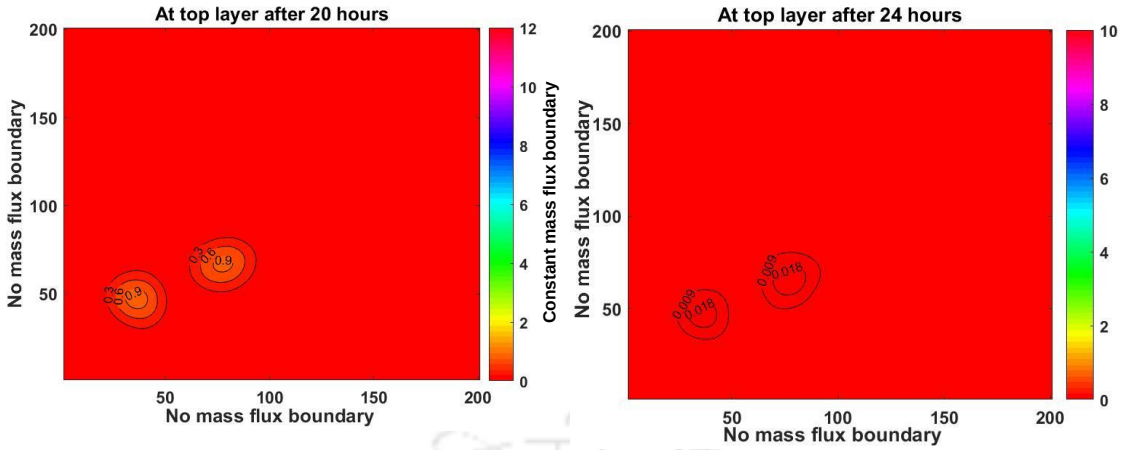
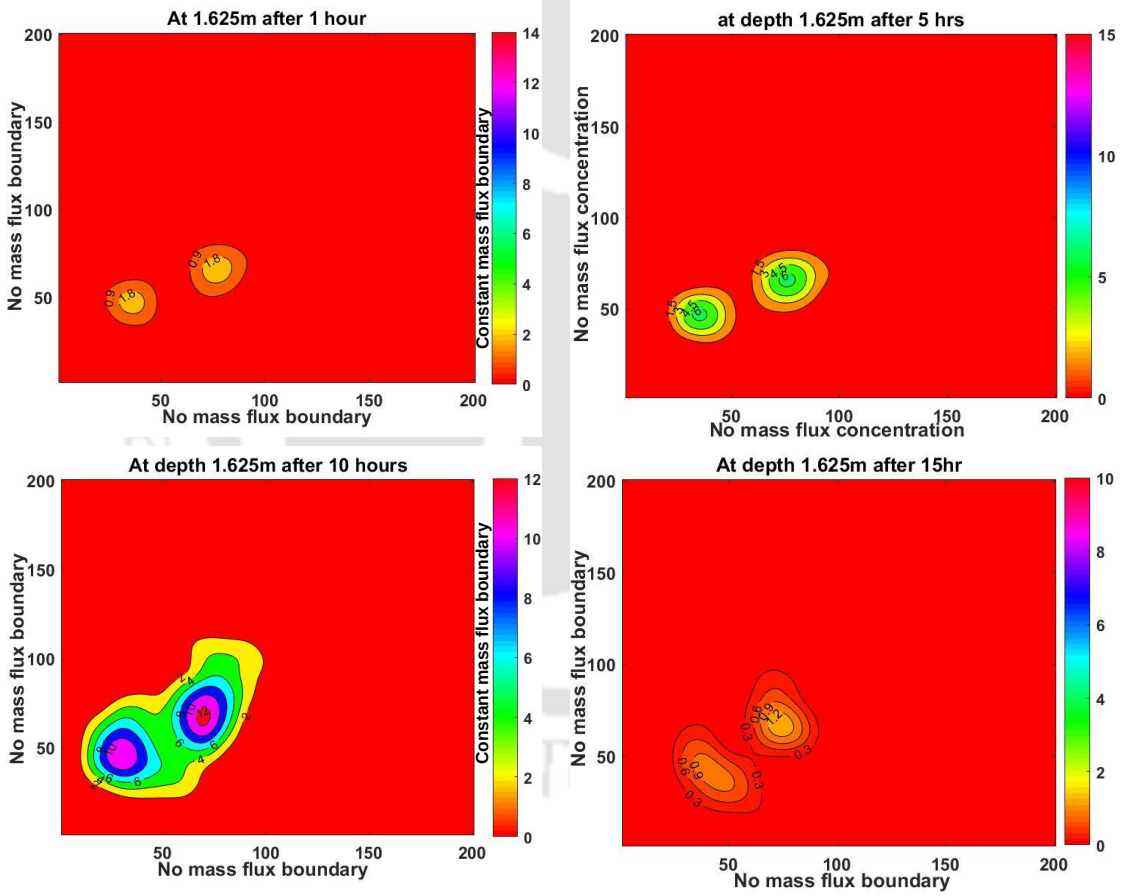


Figure 3. 15: Contour plots showing the spreading of virus concentration (g/h) at different time steps in the top layer (unsaturated zone).



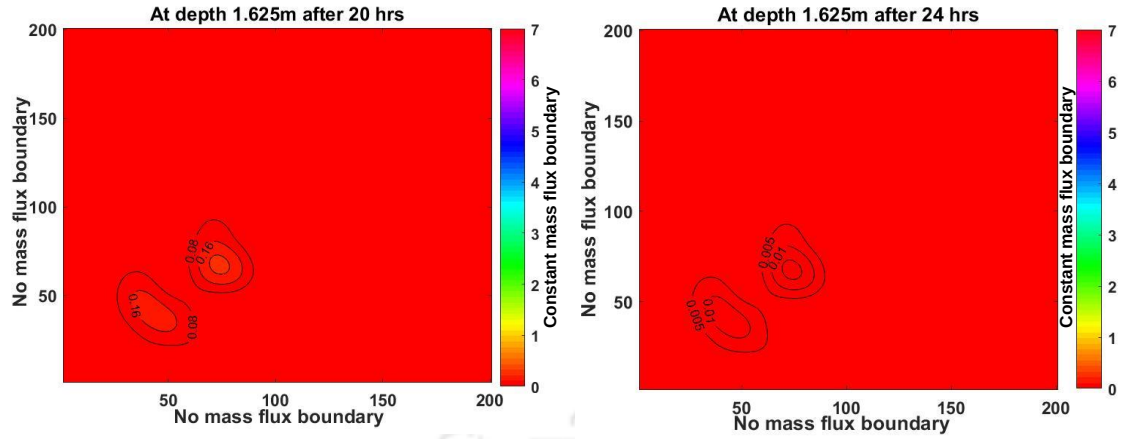


Figure 3. 16: Contour plots showing the spreading of virus concentration (g/h) with time at a depth of 1.625m (saturated zone).

Referring to Figure 3. 15 and Figure 3. 16, at the top layer of the study domain that is in the unsaturated zone, the spreading of virus concentration is less compared to the saturated zone. This is because the unsaturated zone typically has lower water content and lower hydraulic conductivity in homogeneous soils compared to its saturated zones. This results in slower flow velocities of water in the unsaturated zone and thus limits the transport and spreading of viruses. The simulation time of this model is 24 hours, and the virus source is active till 10 hours. A higher value of virus concentration (18 g/h) is observed at 10 hours. After this, the rate and extent of virus concentration decrease and tends to zero after 24 hours. In the saturated zone, two pumping wells are located to pump out water from the groundwater aquifer for the first 15 hours. The pumping action has changed the flow direction and velocity of groundwater and thus influenced the movement and distribution of viruses. This results in a high impact on virus concentration in the surrounding area as shown in Figure 3. 16.

3.5.2 Numerical solution of Three-dimensional groundwater flow and transport model for Scenario 2:

Another synthetic groundwater flow model in three-dimensional space is developed for unconfined aquifers considering unsaturated and saturated zones. The dimensions and boundary conditions are the same as considered in Scenario 1. The only difference in this scenario is that this model considers one source location in an unsaturated zone and another in a saturated zone. Two pumping wells is located, and the water is pumped out from the computational layer, k=8, 9, 10 of the saturated zone. The detail of the virus source flux release and the activation of the pumping well is given in Table 3. 4. The graphical representation of

the study area is shown in Figure 3. 17 and the grid size distribution along with the pumping well location and source location are shown in Figure 3. 18.

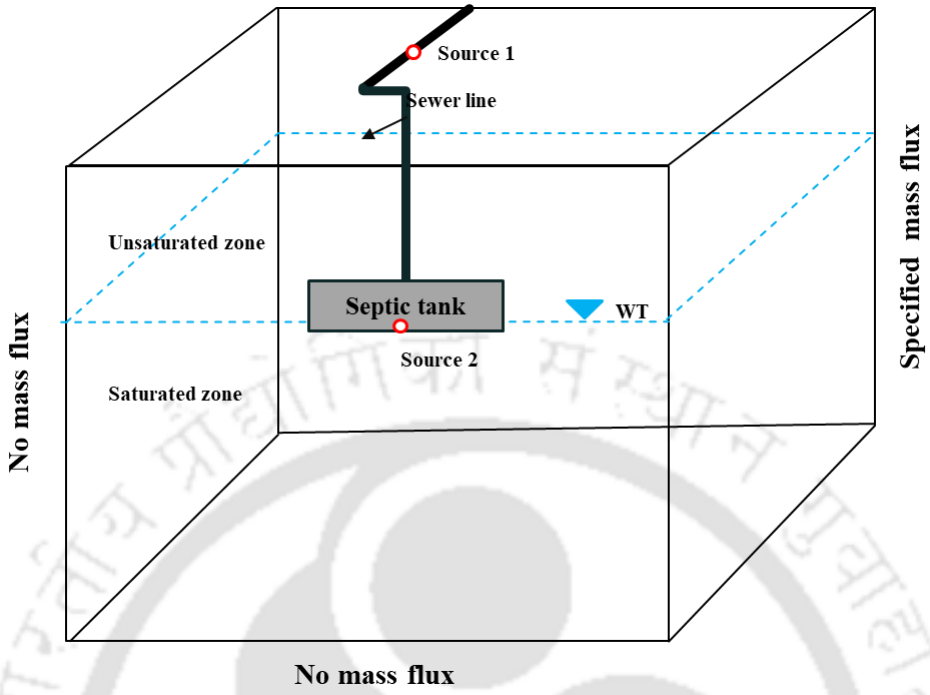
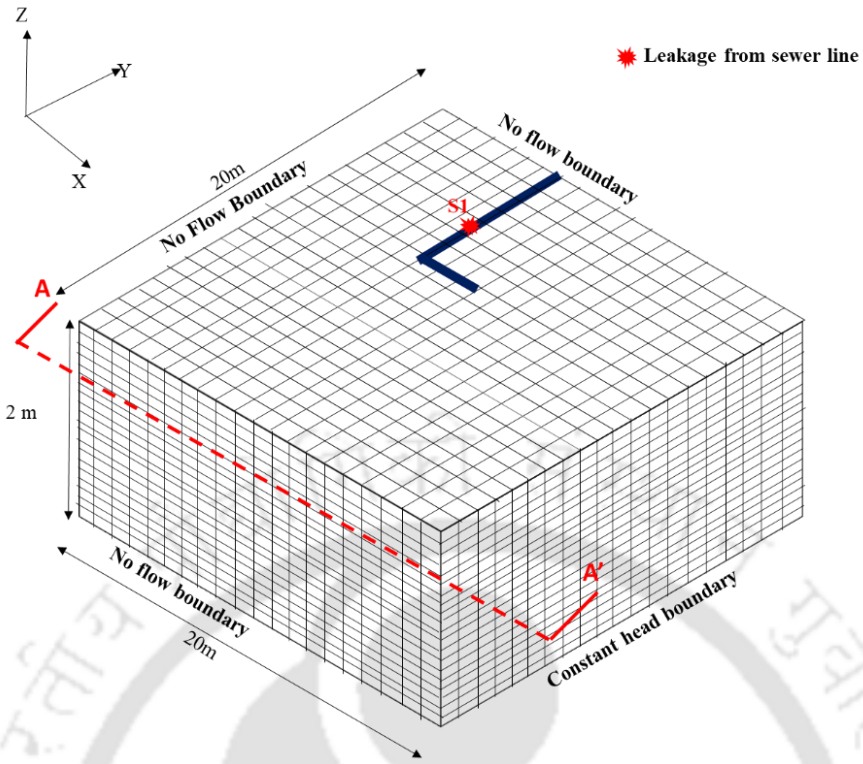


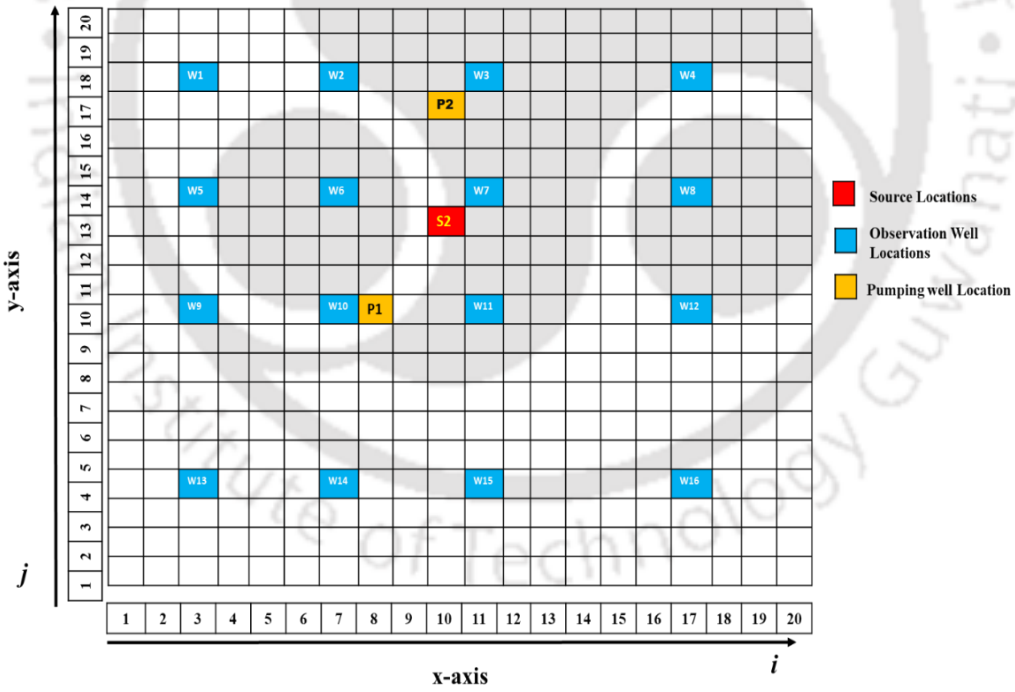
Figure 3. 17: Graphical representation of the scenario 2.

Table 3. 4: Pumping well and source locations along with fluxes for scenario 2

	Input data	Value	Active	Inactive
1	Input source flux(If_1)	0.35 m/h	0-10 hours	10-24 hours
2	Input source flux (If_2)	0.17 m/h	0-5 hours	5-24 hours
3	Pumping rate(q_p)	0.25 m/h	0-15 hours	15-24 hours
4	Source location 1 in x,y,z co-ordinate axis (Sl_1)	(8,16,1)	-	-
5	Source location 2 in x,y,z co-ordinate axis (Sl_2)	(10,13,8)	-	-
6	Source release strengths of location 1 (Sf_1)	300 g/h	0-10 hours	10-24hours
7	Source release strengths of location 2 (Sf_2)	800 g/h	0-5 hours	5-24 hours



(a) Three- dimensional grid showing the source location and boundary condition

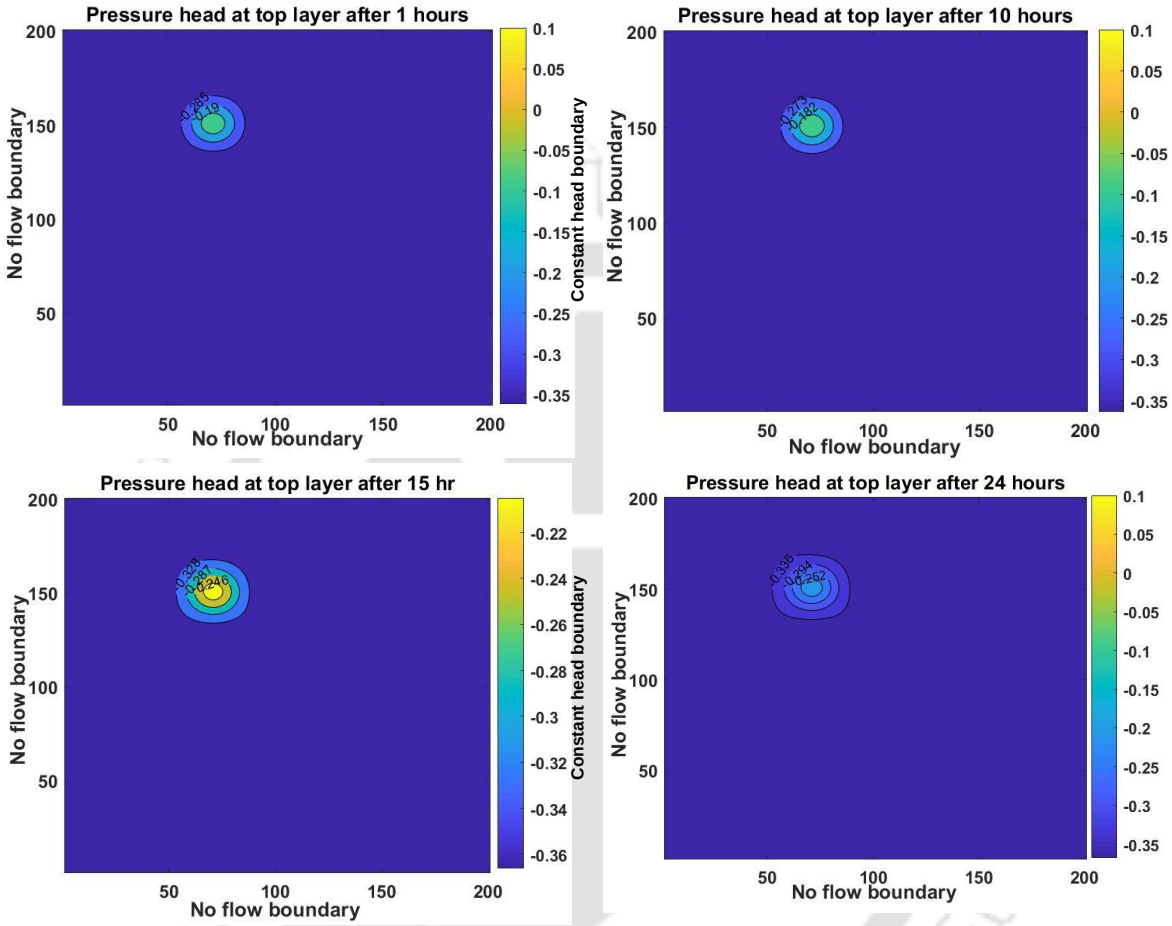


(b) Source location, pumping well location and observation well at layer k =8 (water table position)

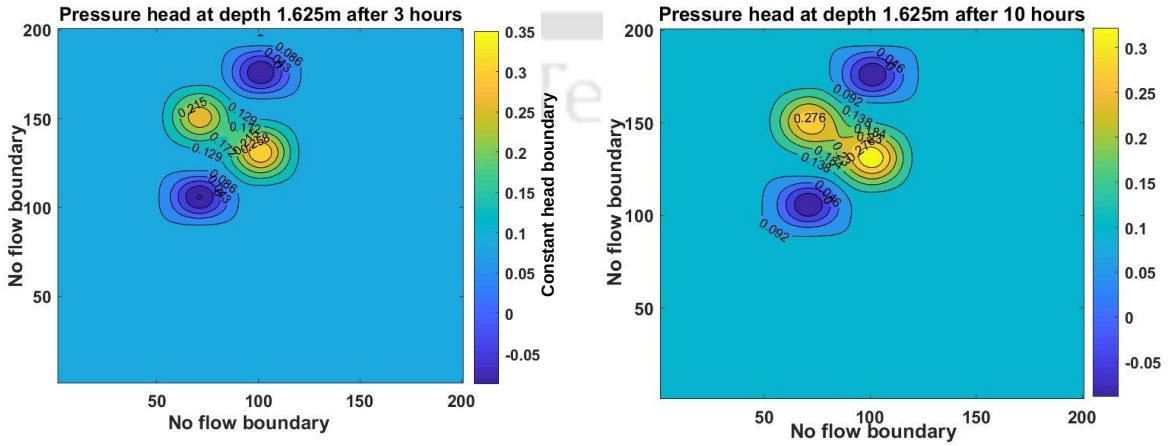
Figure 3. 18: Grid size distribution showing the source location, observation wells and pumping well location considered in Scenario 2.

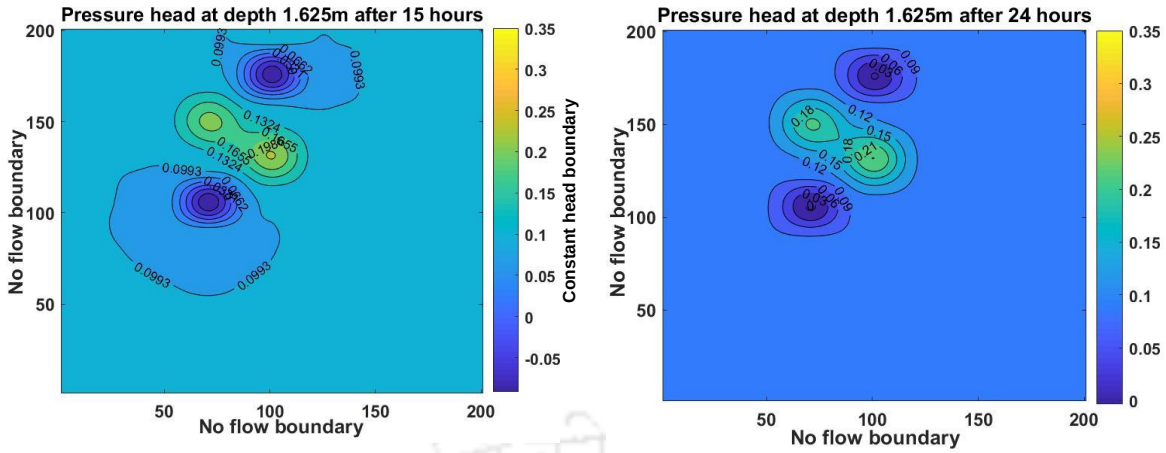
3.5.2.1 Variation of pressure head (m) obtained from the Scenario2:

The groundwater flow model for scenario 2 is simulated for 24 hours, and the pressure head obtained at different time steps is presented using contour plots, as shown in Figure 3. 19. This provides a visual representation of the distribution of pressure head values in the unsaturated and saturated zones of the proposed study area.



(a) Pressure head (m) in unsaturated zone





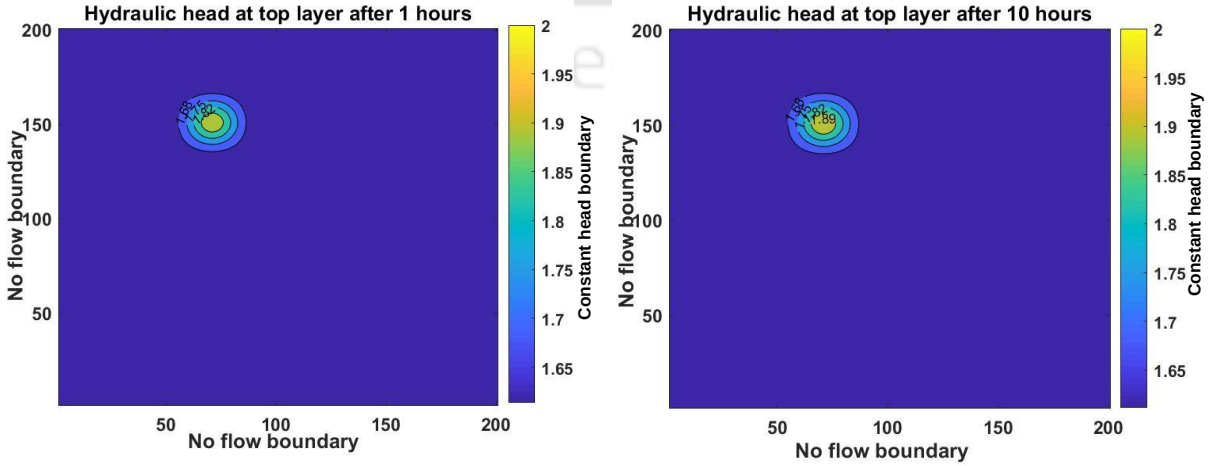
(b) Pressure head (m) in saturated zone

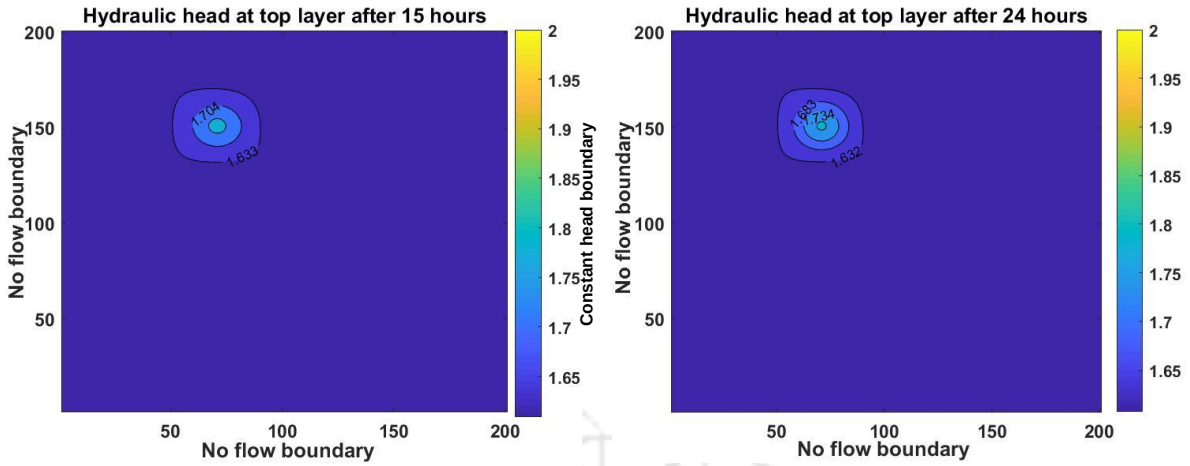
Figure 3. 19: Contour plots showing the pressure head at different time steps in the saturated and unsaturated zone in scenario 2.

Figure 3. 15 shows that one source is present in the top layer, so a negative pressure head is observed over the entire layer near the source location. This indicates that the aquifer at the top layer is unsaturated or partially saturated. After simulation, the pressure head obtained at 10 hours shows a positive pressure head in the source location due to recharge (see Figure 3. 19(a)). This indicates that the pore spaces within the region are filled with water. In this scenario, two pumping wells and one virus source are present in the saturated zone. While referring to the pressure head distribution Figure 3. 19 (b), a negative pressure head is seen in the pumping well locations, and an increase in pressure head is seen in the virus source locations.

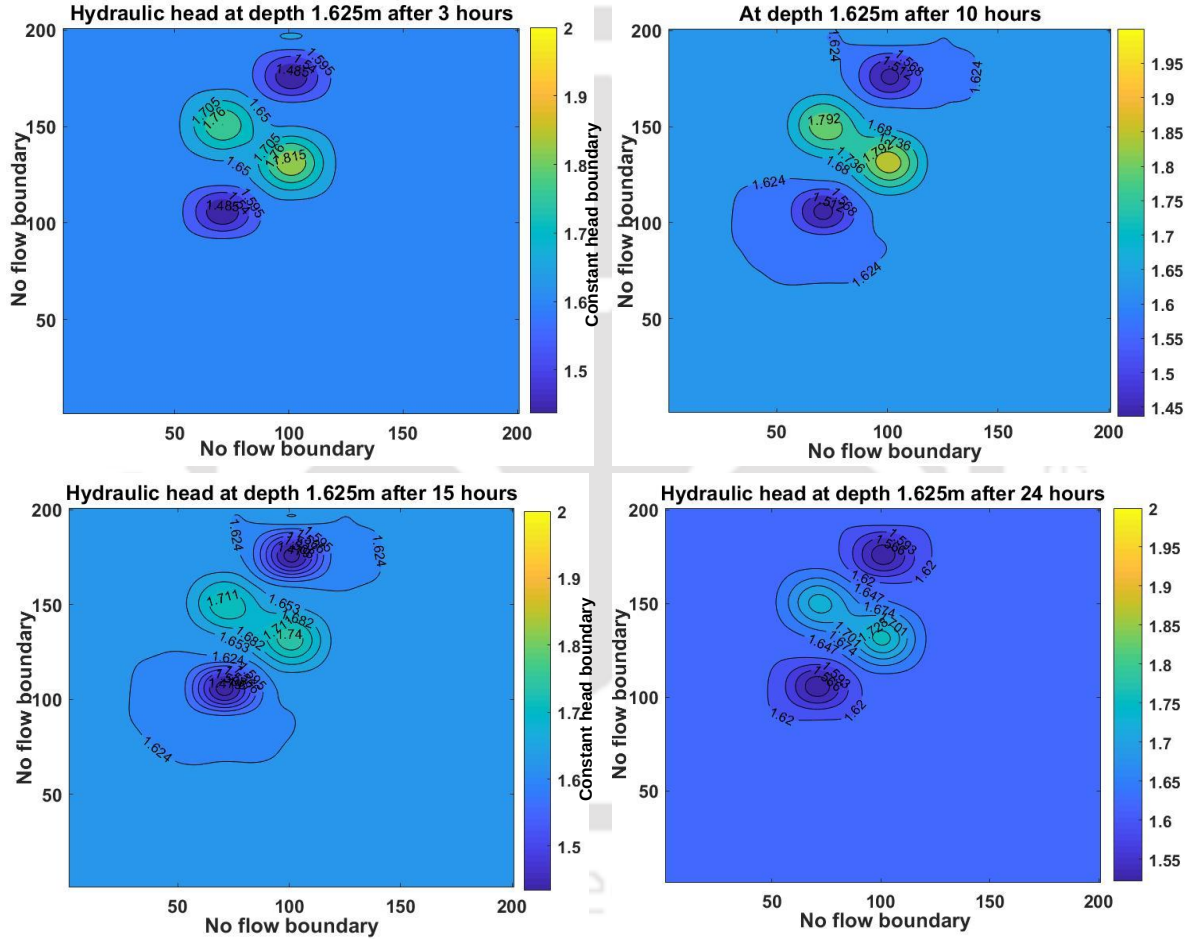
3.5.2.2 Variation of hydraulic head (m) obtained from the Scenario 2:

As mentioned in scenario 1, this model also provides the spatial and temporal variation of hydraulic heads within the study domain. Figure 3. 20 provides the contour plots of the hydraulic head in both saturated and unsaturated zones with time.





(a) Hydraulic head (m) in unsaturated zone



(b) Hydraulic head (m) in saturated zone

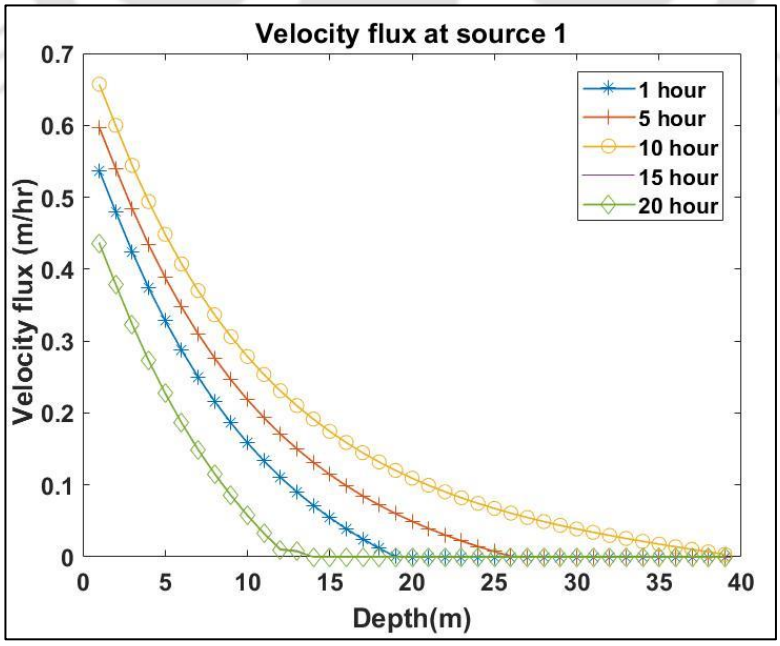
Figure 3. 20: contour plots showing the hydraulic head(m) at different time steps in saturated and unsaturated zones in scenario 2.

Initially, the water table was located at an elevation of 1.625m with respect to the datum, but after 15 hours of pumping, the hydraulic head has lowered to 1.475 m elevation in pumping well locations. This is because the pumping reduces pressure and draws water into the well. On the other hand, the input flux recharge in the source location adds moisture to the aquifer,

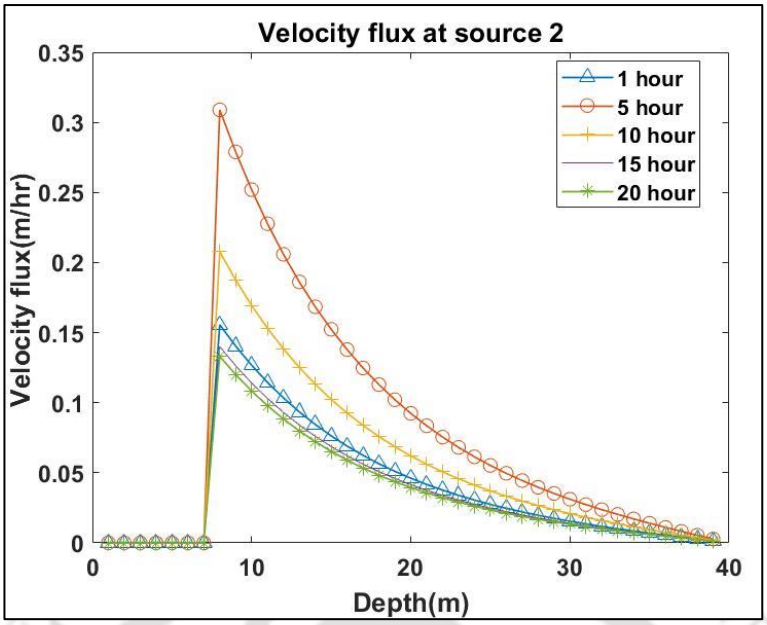
which increases the water table elevation and subsequently raises the hydraulic head. It is found that the hydraulic head has raised to 1.89m from the datum at the top layer where source 1 is present. In the saturated zone that is at a depth of 1.625m from the datum, the hydraulic head has raised up to 1.825m in 10 hours. After that, the hydraulic head at both locations eventually decreased, and can be clearly seen in Figure 3. 20.

3.5.2.3 Velocity vectors obtained from the scenario 2:

In Scenario 2, the velocity flux distribution is presented to depict both the spatial and temporal variations at the two source locations. This distribution provides insights into how the velocity flux values change over time and across different locations within the aquifer. The plot of the velocity flux distribution in Scenario 2 is presented in Figure 3. 21. From the velocity flux plot, it is observed that in source 1, velocity flux increases sources during the first 10 hours of the time period. Whereas in source 2, the flux increases during the first 5 hours. This is because the input flux recharge in source 2 is for 5 hours, and for source 1 it is for 10 hours. Figure 3.18 shows that the velocity has increased drastically after a certain depth. This is because at that particular depth, the source is present, and water is getting recharged at that position. After the initial increase, the velocity flux decreases until it reaches a lower value at 15 hours. This indicates a gradual reduction in groundwater flow rate in both sources during this time interval. Beyond 15 hours, the velocity flux remains constant. This implies that the groundwater flow rate from both sources stabilizes, and there is no further change in the velocity flux values. The effect of velocity flux variation is seen in virus concentration.



(a)

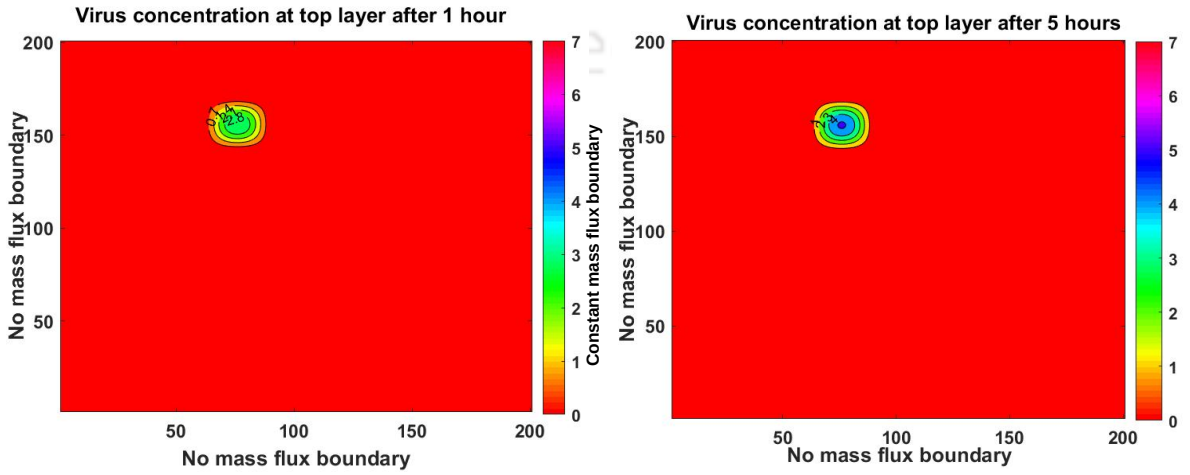


(b)

Figure 3. 21: Variation of velocity flux at two source locations with depth at different time steps in scenario 2.

3.5.2.4 Variation of virus concentration obtained from the scenario 2:

In this scenario, sources of virus contamination are present in two locations. The first location is in the unsaturated zone, and the second is in the saturated zone. The model is simulated for 24 hours, and the source is kept active for the first 10 hours in source 1, and in source 2 it is kept active for the first 5 hours; after that, it is marked inactive. After 24 hours of simulation, the spatial and temporal variation of virus concentration is generated and is presented in contour plots, as shown in Figure 3. 22 and Figure 3. 23. The data set presented in the contour plots is obtained after interpolating the data generated from the simulation model. The x and y axes of the plot represent the spatial coordinates, while the contour lines represent different virus concentration levels in g/h.



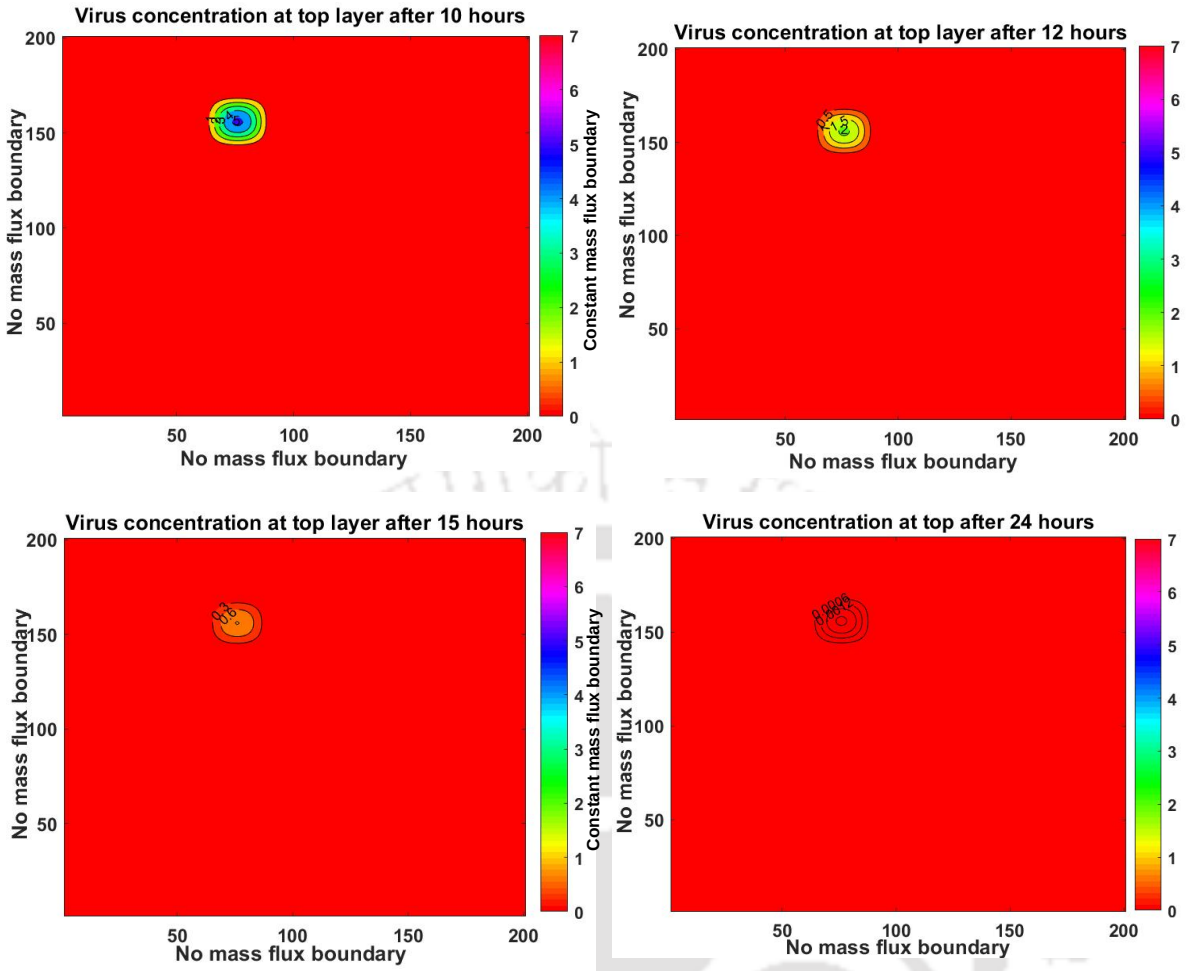
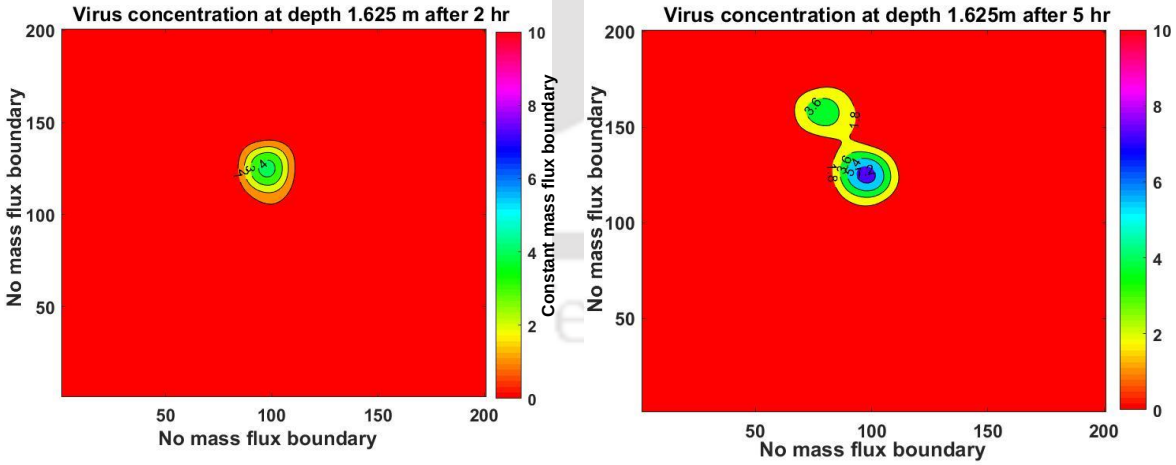


Figure 3. 22: Contour plots showing the spreading of virus concentration (g/hr) at different time steps in the top layer (unsaturated zone) in scenario 2.



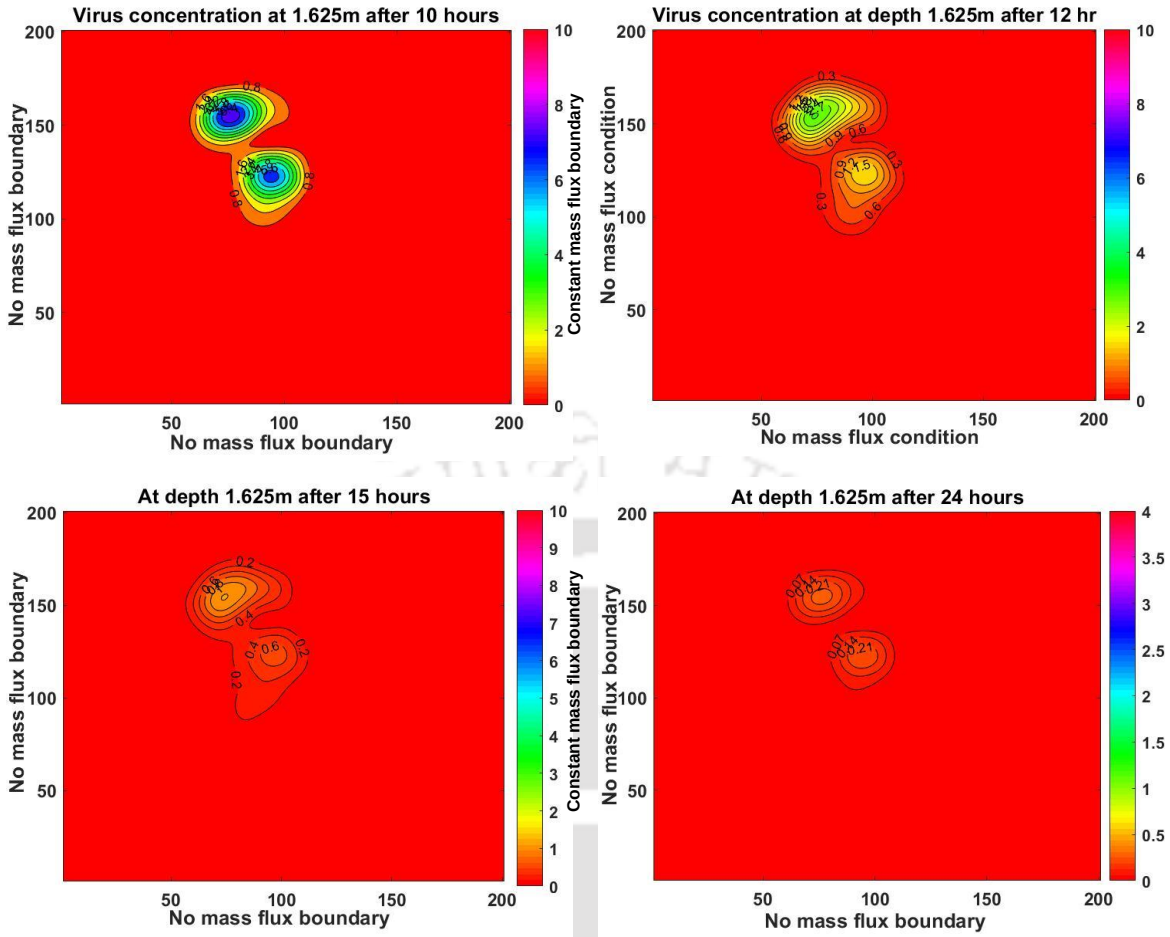


Figure 3. 23: Contour plots showing the spreading of virus concentration with time at depth of 1.625m (saturated zone) for scenario 2

Observation: In the unsaturated zone, as shown in Figure 3. 22, it is observed that the spreading of virus concentration is less due to lower water content and lower hydraulic conductivity. A higher value of virus concentration (8.2 g/h) is observed at 10 hours; after this, the rate and extent of virus concentration decrease and tend to zero after 24 hours. While referring to Figure 3. 23, it is observed that at 2 hours, the virus contamination due to source 1 has not reached the groundwater table. However, after 5 hours, the contour of virus concentration due to source 1 is seen in the saturated zone, meaning that the virus took around 5 hours to reach the groundwater. In Source 2, the value of virus concentration is found to be high at 5 hours; after that, the contamination strength has lowered. In the saturated zone, two pumping wells are located to pump out water from the groundwater for the first 15 hours. The pumping effect is observed in the movement and distribution of viruses in groundwater.

3.6 Conclusions

A numerical simulation model is indeed important for any source identification and management policy development in groundwater aquifers. It provides a quantitative tool to understand and predict the behavior of groundwater flow and contaminant transport. Existing models like HYDRUS-3D and MT3DMS are commonly used but have limitations, such as MT3DMS's inability to account for variable saturation degrees and HYDRUS-3D's challenges with variations in activation and inactivation rates of viruses.

- In conclusion, the three-dimensional virus transport model developed for unconfined aquifers overcomes limitations of existing models, providing a comprehensive tool for understanding and managing groundwater contamination. The model is coded in MATLAB environment
- The model incorporates variations in input parameters based on saturation degrees, providing a more realistic representation of virus transport in porous media.
- Calibration and validation against field and laboratory data ensure the model's accuracy and reliability in predicting virus fate and transport processes in the subsurface environment.
- Overall, the developed three-dimensional virus transport model offers valuable insights into the fate and transport of viruses in the subsurface environment.
- The model currently assumes homogeneous soil profiles and is limited to two saturation levels due to data constraints. To enhance accuracy and applicability, future improvements may involve incorporating virus parameters for diverse saturation levels and expanding the model to account for heterogeneous soil profiles.

CHAPTER 4

PARAMETER ESTIMATION AND GLOBAL SENSITIVITY ANALYSIS OF FLOW AND VIRUS TRANSPORT PARAMETERS

4.1 Overview:

This chapter provides two parameter estimation models for unconfined groundwater aquifers. The first model focuses on estimating flow parameters, while the second model is designed to estimate virus transport parameters. Both models utilize inverse optimization-based techniques to infer the parameters based on available data. After parameter estimation, global sensitivity analysis is performed to assess the sensitivity of the model output to each estimated parameter. This helps researchers and practitioners identify the most influential parameters and prioritize their efforts to improve their accuracy.

4.2 Introduction:

The flow and virus transport models for unconfined groundwater aquifers can be developed for unsaturated and saturated zones. These models incorporate various soil and hydraulic parameters influencing groundwater's movement and viruses' transport within the aquifer. However, in certain situations, it can be challenging to directly measure these parameters with the desired level of accuracy in the field or laboratory scale. In such cases, inverse optimization-based parameter estimation methods can be employed to estimate the unknown groundwater flow and transport parameters. This study proposes two models to predict flow and virus transport parameters in unconfined groundwater aquifers that consider unsaturated and saturated zones. The existing models available in the literature are not designed to simultaneously estimate the parameters for both the unsaturated and saturated zones in a coupled manner (Huang & Dun, 2008; Şahin, 2018;Samuel & Jha, 2003 ;Swathi & Eldho, 2018; Liu et al., 2020; Eryiğit, M. 2021). In the first model, a computationally efficient parameter estimation model is developed to simultaneously estimate the saturated and unsaturated flow parameters. As such, an approximate groundwater simulation model is developed using Artificial Neural Network (ANN). This ANN model is coupled with the optimization model to satisfy the objective function by minimizing the error between the observed and simulated hydraulic heads. The virus transport parameters are estimated in the second model since virus transport models have traditionally relied on the constant detachment

model to estimate transport parameters (Schijven & Šimůnek, 2002; Ratha et al., 2009; Ojha et al., 2012; Bhattacharjya et al.,2015). However, this approach may not always be accurate since the detachment-attachment rates of viruses are known to vary based on the saturation level of the medium. The study, therefore attempted to developed a parameter estimation model by considering variable attachment detachment rates for saturation levels. The transport parameters estimated are - the distribution coefficient, the inactivation rate, and the adsorption rate of viruses at soil-water and air-water interfaces. Furthermore, the parameter estimation problem is considered as a nonconvex problem, meaning that multiple solutions can satisfy the imposed constraints, resulting in the non-uniqueness of the solutions. To tackle such problems, meta-heuristic algorithms are employed in both models, as they have proven effective in solving such nonconvex problems. Additionally, sensitivity analysis is conducted to examine the relative importance of each flow and virus transport parameter to the model output. This helps to identify the parameters that significantly uncertain and must be characterized more accurately.

4.3 Estimation of flow parameters using ANN and SFLA (Model 1):

The study utilises the constitutive relationship developed by Van Genuchten, M. Th. & Nielsen, (1985) that relates unsaturated hydraulic properties like pressure head (h), water content (θ), and hydraulic conductivity (K), mutually. To define the constitutive relations accurately, five hydraulic parameters need to be estimated: saturated water content (θs), residual water content (θr), soil parameter (α), saturated hydraulic conductivity (Ks), and fitting emperical parameter (n). These parameters capture the behavior of water movement in the unsaturated zone of the aquifer. In addition to these unsaturated flow parameters, the specific storage (Ss) is an essential parameter that characterizes the saturated zone of the aquifer. Specific storage represents the ability of the aquifer to store and release water under changing hydraulic conditions. Therefore, in this study, the optimization model considers six decision variables: θs, θr, α, Ks, n, and Ss. These parameters collectively represent the flow properties of the groundwater aquifer's saturated and unsaturated zones.

4.3.1 Methodology

In the flow parameter estimation process, an inverse optimization technique is employed. The goal is to minimize the error function between the observed and simulated hydraulic heads. The observed head is obtained from field observation wells, and the groundwater simulation

model generates the simulated head. The optimization algorithm used in this model is the Shuffled Frog leaping Algorithm (SFLA). Initially, candidate solutions are generated by the optimization algorithm, which serves as input parameters for the numerical groundwater simulation model. The simulation model uses these candidate solutions to calculate the aquifer's spatial and temporal hydraulic head distribution. The optimization algorithm modifies or updates the candidate solutions based on the objective function value. This process is repeated iteratively, with the simulation model generating new hydraulic head distributions based on the updated candidate solutions and the objective function value being re-evaluated. The iterative optimization continues until the optimal solution is achieved, corresponding to the best fit between the simulated and observed hydraulic heads. The objective function used to estimate all the flow parameters as given by Eryiğit (2021) is shown equation (4.1).

$$\text{Minimize} \quad f_x = \sum_{j=1}^{M}\sum_{i=1}^{N}\left|OH_i^j - SH_i^j\right| \tag{4.1}$$

$$\text{Subject to} \quad H = f\left(\theta_s, \theta_r, \alpha, n, K_s, S_s\right)$$

$$\theta_s^{\min} \le \theta_s \le \theta_s^{\max}$$

$$\theta_r^{\min} \le \theta_r \le \theta_r^{\max}$$

$$\alpha^{\min} \le \alpha \le \alpha^{\max}$$

$$n^{\min} \le n \le n^{\max}$$

$$K_s^{\min} \le K_s \le K_s^{\max}$$

$$S_s^{\min} \le S_s \le S_s^{\max}$$

Where $\boldsymbol{f_x}$ is the objective function for the present optimization model with $\boldsymbol{x}$ number of parameters; OH_i^j is the observed hydraulic head at $\boldsymbol{i^{th}}$ time step for $\boldsymbol{j^{th}}$ well location; SH_i^j is the simulated hydraulic head at $\boldsymbol{i^{th}}$ time step for $\boldsymbol{j^{th}}$ well location obtained from numerical simulation model; $\boldsymbol{M}$ is the total number of observation wells and; N is the total number of time steps; The suffix '*min*' and '*max*' represents the lower bound and upper bound of the variables - $\theta_s, \theta_r, \alpha, n, K_s, S_s$.

When the numerical simulation flow model is linked with the optimization model to estimate the parameters, this combination takes around one day, 5 hours, 45 minutes, and 10 seconds, which is very time-consuming. This model was run in computer system with specification of processor: Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz 3.41 GHz, 8GB RAM. Thus, an alternative simulator addresses the time-consuming parameter estimation process by replacing

the numerical simulation flow model with an Artificial Neural Network (ANN). In this study, ANN models are developed, taking the six decision variables (θs, θr, α, Ks, n, and Ss) as inputs and the hydraulic head at different time steps as outputs for various observation wells. The flow parameter estimation model is developed by integrating the ANN simulator with the optimization model. The estimation is completed in 4 hours, 57 minutes, and 12 seconds, considerably less than the previous parameter estimation model. The steps involved in developing the ANN model, including data preprocessing, network architecture design, training, and validation, are explained in section 4.3.2. Figure 4.1 shows the flowchart of the flow parameter estimation model using ANN-SFLA.

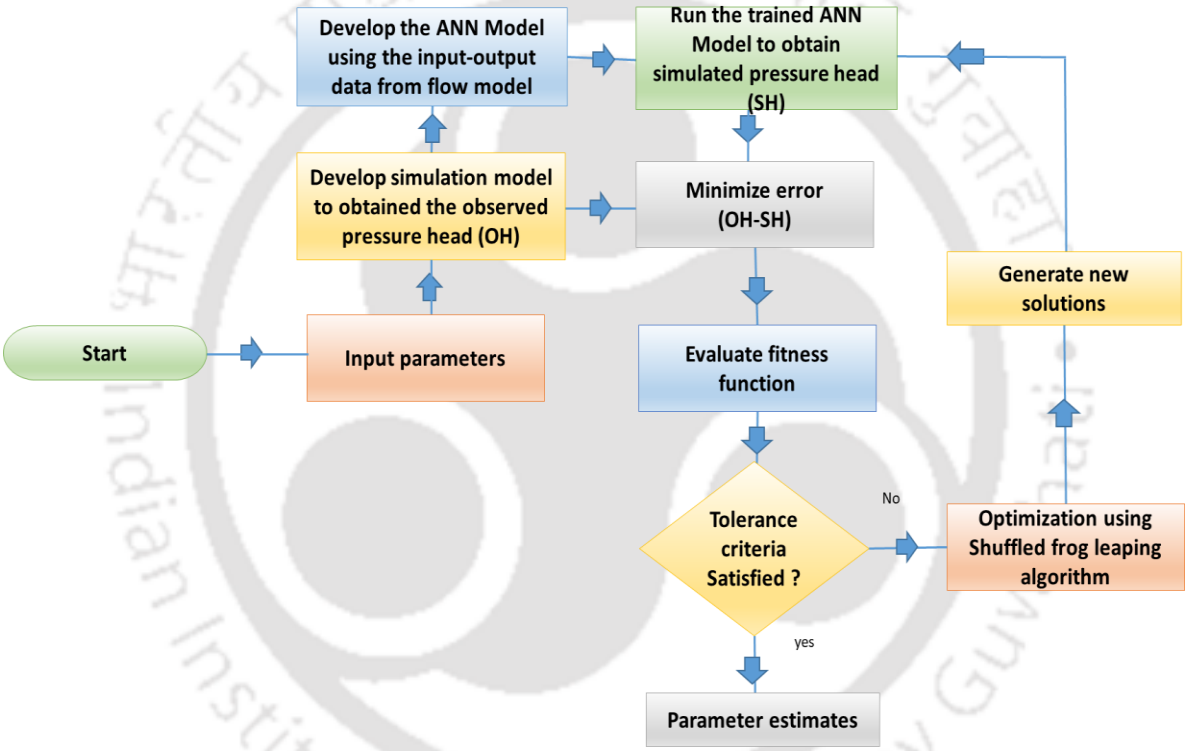


Figure 4.1: Flow chart showing the ANN-SFLA parameter estimation model.

4.3.2 Development of approximate simulator using ANN:

Artificial neural networks (ANNs) have been proven to be an effective and popular alternative to numerical aquifer simulations (Afzaal et al., 2020; Chen et al., 2019; El Amri et al., 2022; Mohanty et al., 2013 ; Nikolos et al.,2008 & 2010; Papadopoulou et al., 2010; Goh et al., 2020; Liu et al., 2022; Zhang et al., 2019).The proposed methodology uses the ANN model as the surrogate groundwater flow model for the particular study area. The ANN model is trained using the data generated from the groundwater flow model developed in Chapter 3. The input to the ANN model is the groundwater flow parameters. The output from the ANN model is the

hydraulic head at different observation well for different time steps. A total of six ANN models were developed based on the data from six observation wells in this study area. The location of the observation wells is shown in Figure 4.2.

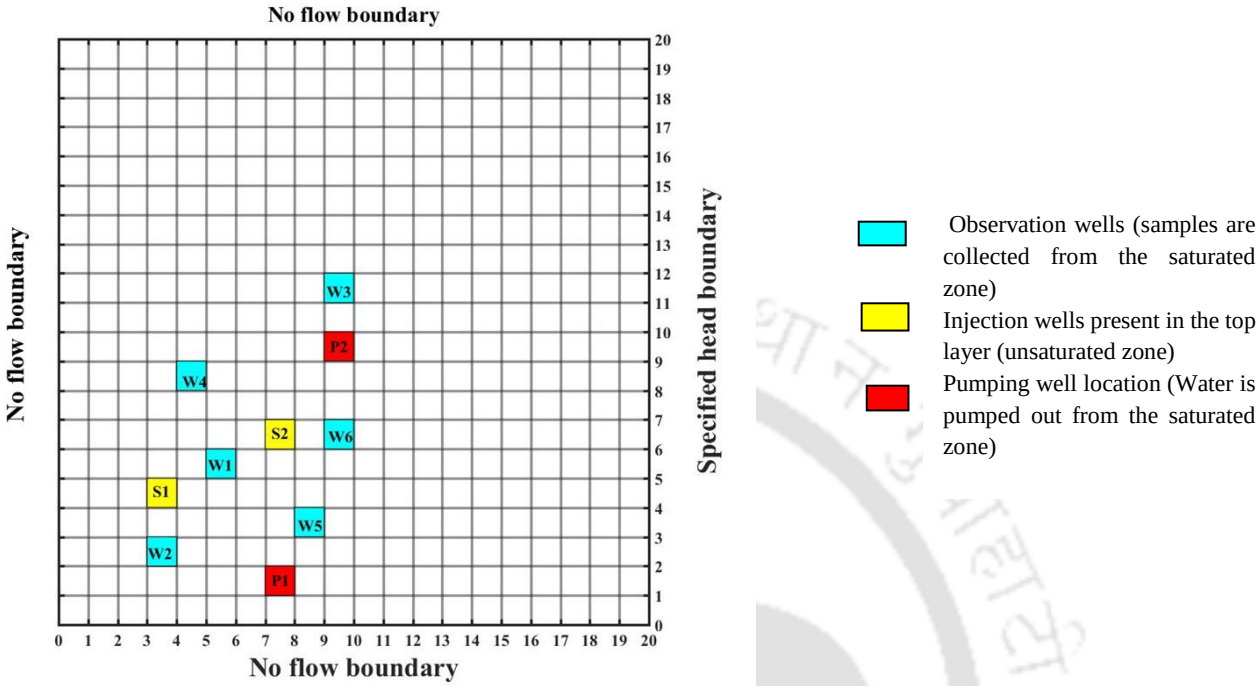


Figure 4.2: Grid distribution of the flow model showing the observation well locations, injection wells, and the pumping well in XY plane.

4.3.2.1 Generation of ANN pattern

The current methodology employs a single hidden layer architecture. A decision regarding the number of hidden layers and neurons in the architecture is made through trial and error. An ANN model undergoes training, testing, and validating processes during its development process. The ANN model is trained using the Levenberg-Marquardt algorithm. A linear transfer function is used in the output layer, while a unipolar sigmoidal transfer function is used in the hidden layer. Using the groundwater flow model developed in the previous chapter, 1000 input-output patterns data were generated. Of these, 60% of the generated data is used for training the ANN model, which can produce desirable outputs based on the input pattern. Further, 40% of the patterns are used to test and validate the ANN models.

4.3.2.2 Architecture of the ANN model for the study area

In this present study, six flow parameters are proposed to be estimated in the groundwater flow model. As such, a total number of six parameters are used as the input to the ANN model. The numerical simulation flow model is simulated for 24 hours with a time interval of 30 minutes. As such, the output from the ANN model is for 48 time steps. Six observation wells are selected

in the unconfined aquifer, resulting in six ANN models. This ANN model results in a three-layer ANN model: one input layer, one hidden layer, and one output layer. A schematic representation of the ANN model with a 6-40-48 neuron pattern is shown in Figure 4. 3.

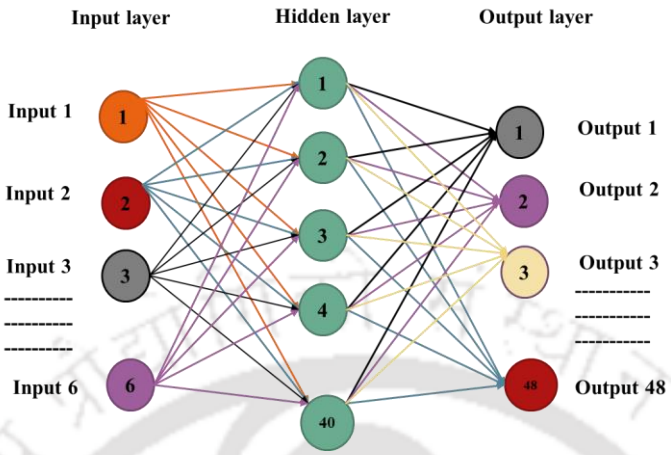


Figure 4. 3: ANN model network

4.3.2.3 Transfer function and optimization algorithm adopted for the ANN model

Many transfer functions are available in MATLAB, such as *tansig*, *logsig*, *purelin*, etc. Figure 4. 4 shows the most commonly adopted transfer functions the backpropagation networks use. However, combining *tansig-purelin* function for hidden and output layers produced the best mean square error (MSE), computation time, and iteration time (Borah and Bhattacharjya, 2014). Based on the referred report, the *tansig* and *purelin* transfer functions were used in developing the ANN model. The *tansig* function generates an output ranging from -1 to +1 and is a very effective method when speed is considered an important criterion. Sometimes, models possess linear characteristics and require minimal parameters. Whenever these situations arise, *purelin* transfer functions can produce any value.

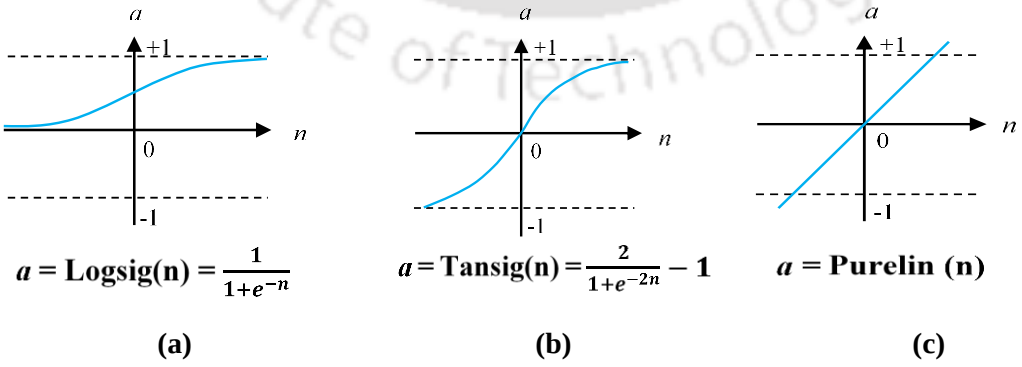


Figure 4. 4:Transfer function used in the feed-forward network (a) Log-Sigmoid (b) Tan-Sigmoid, and (c) Purelin.

After selecting an ANN network, the next important step is to select an optimization algorithm. In order to train the current ANN model, the *trainlm* function in MATLAB has been used. The weight and bias are updated using the principle Levenberg-Marquardt algorithm by the *trainlm* function. Backward propagation is recommended as the first choice for supervised algorithms since it is considered the fastest (Gaur et al., 2013; Noori et al., 2010; Sharma et al., 2008; A. Singh et al., 2022; Siregar & Wanto, 2017).The combined transfer functions and the *trainlm* function were used to train the network due to their computational efficiency and good convergence power.

4.3.2.4 Evaluation of ANN model performance using statistical criteria

Three statistical criteria is adopted to evaluate the ANN model's performance. This includes the Average Absolute Relative Error (AARE), Root Mean Square Error (RMSE), and the Coefficient of Correlation (R). The following section explains the mathematical expressions for all three statistical criteria.

4.3.2.4.1 Average Absolute Relative Error (AARE)

In simple terms, AARE represents the average difference between observed and simulated hydraulic head divided by simulated hydraulic heads. The following expression can be used to calculate AARE:

$$AARE = \frac{1}{n_c}\sum_{n_c=1}^{n_c}\left|\frac{OH_i^j - SH_i^j}{SH_i^j}\right| \times 100 \tag{4.2}$$

4.3.2.4.2 Root Mean Square Error (RMSE)

A root mean square error is defined as the average value of the squares of the difference between simulated and observed concentrations. RMSE can be calculated using the equation as

$$RMSE = \sqrt{\frac{1}{n_c}\sum_{n_c=1}^{n_c}\left(OH_i^j - SH_i^j\right)^2} \tag{4.3}$$

4.3.2.4.3 Coefficient of Correlation (R)

The coefficient of correlation is the sum of the products of the deviation of each data from its respective mean, divided by the product of the number of sets and the standard deviations. It can be expressed as

$$R = \frac{\sum_{n_c=1}^{n_c}\left(OH_i^j - \overline{OH_i^j}\right)\left(SH_i^j - \overline{SH_i^j}\right)}{\sqrt{\sum_{n_c=1}^{n_c}\left(OH_i^j - \overline{OH_i^j}\right)^2\left(SH_i^j - \overline{SH_i^j}\right)^2}} \tag{4.4}$$

Here, the observed hydraulic head (OH) is obtained from the numerical simulation model and the simulated hydraulic head is obtained from the ANN model. Lower values of AARE and RMSE signifies a good performance of the ANN model. If the value of R^2 is close to 1, then there is a close relationship between the observed and the simulated hydraulic head.

4.3.3 Shuffled Frog Leaping Algorithm

The shuffled frog leaping algorithm (SFLA) is used as the optimization algorithm to estimate flow parameters in this study. The shuffled frog leaping algorithm (SFLA) is a nature-inspired swarm-based metaheuristic optimization algorithm inspired by the frog's behavior in finding food in wetlands. The algorithm is a combination of Particle Swarm Optimization and a Memetic algorithm. As such, it is simple, fast, and efficient global optimizer. It involves interactive individuals (frogs) carrying memes and exchanging information among themselves globally. As the flow parameter estimation is a nonconvex problem having multiple local optima, the SFLA algorithm is suitably employed. The steps involved in SFLA are depicted in Figure 4.5 and are coded in a MATLAB environment.

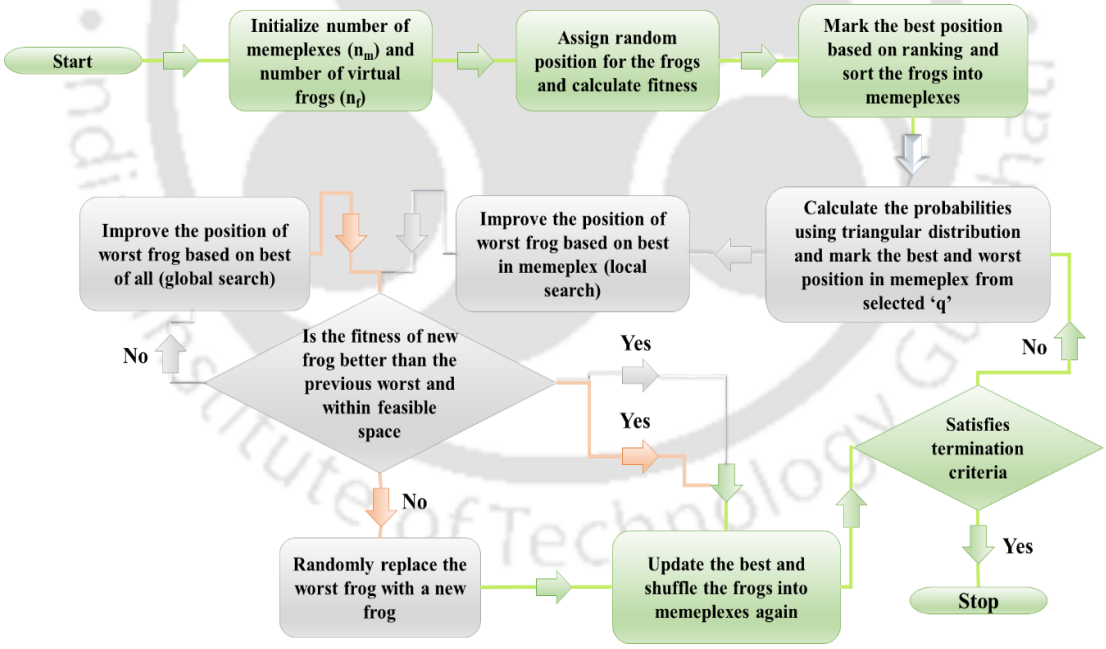


Figure 4.5: Flowchart showing the process of shuffled frog leaping algorithm (SFLA)

The steps of the shuffled frog-leaping algorithm are as follows:

Step 1: To begin the problem, the number of memeplexes (n_m) and virtual frogs within each memeplex (n_f) are taken at the beginning. This gives the total number of frogs as $n_m \times n_f$.

Step 2: The algorithm continues by assigning a random position to all the frogs and calculating the corresponding fitness.

Step 3: The best frog is then marked as the global best, and then the frogs are sorted into memeplexes. This portion of the flowchart is presented in green color. The next portion is local evolution in each memeplex, represented in grey colour as shown in Figure 4.5.

Step 4: The frogs are distributed using the triangular distribution, as shown in equation (4.5), where p_i represents the probability for triangular distribution, and from them, q_x frogs are selected.

$$p_j = \frac{2\left(n_f + 1 - j\right)}{n_f\left(n_f + 1\right)}, \text{ where } j = 1 \ldots . n_f \tag{4.5}$$

The best and the worst frogs are then marked. The position of the worst frog is improved by choosing an appropriate step size based on the best of the memeplex.

Step 5: The condition of whether this frog is better than the previous worst and within the feasible space is checked. If the condition is satisfied, then the step to update the memeplex and shuffling is continued. If the condition is not satisfied, then the worst position is improved based on the global best. This set of conditions is represented with blue connectors and arrows with the decision matrix in the flowchart.

Step 6: The improved position is again checked with the same condition as the previous one. If the condition is satisfied, then the updating of memeplex with shuffling is continued. If the condition is not satisfied, the new position is improved randomly, and then the memeplex is updated. The flowchart represents this set of conditions with red connectors and arrows.

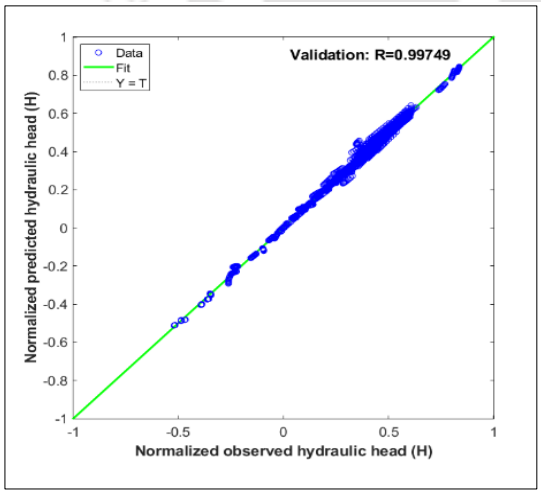
Step 7: Again, the termination criteria are checked after updating the global best, and reshuffling into the memeplexes is done. The global optimal solution is reached if the termination criteria are satisfied. If the termination criteria are not satisfied, then the algorithm resumes the step of evolution from each memeplex.

4.3.4 Results and Discussion (MODEL 1):

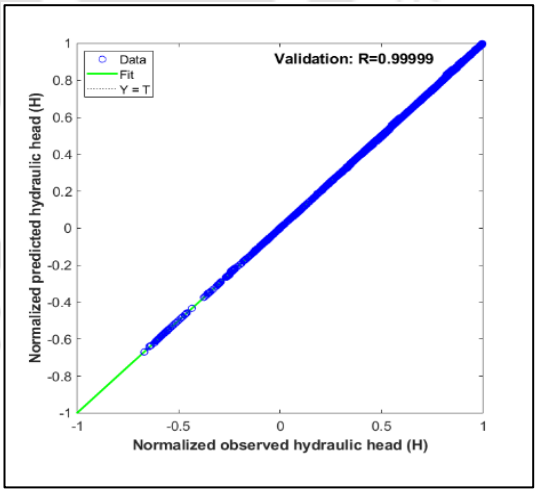
4.3.4.1 Performance of the ANN models:

Using the groundwater simulation model developed in Chapter 3, 1000 input-output patterns are generated. Out of which 60% data is used to train the ANN model and the remaining 40%

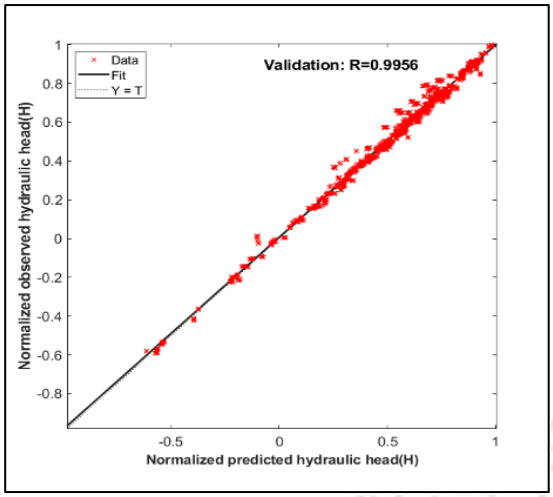
is used for testing and validation. The performance of the developed six ANN models is shown in Figure 4. 6 and Figure 4. 7. Figure 4. 6 illustrates the performance of the six ANN models through scatter plots for each observation well at forty-eight time steps. The coefficient of correlation (R^2) values for the best and worst-performing models are reported as 0.9999 and 0.9375, respectively. A value of R^2 close to 1 indicates a strong correlation between the predicted hydraulic head and the actual hydraulic head, suggesting that the ANN models accurately capture the relationship between the input parameters and the output. Figure 4. 7 shows the performance of all six ANN models in terms of Mean Square Error (MSE) for training, testing, and validating data. The MSE measures the average squared difference between the predicted and actual values. As the models are trained, and the number of epochs increases, the MSE gradually decreases, converging toward the best possible value. The calculated MSE values range from 2.3×10^{-2} to 4.07×10^{-6}, indicating negligible errors. These low MSE values further support the conclusion that the developed ANN models effectively simulate the groundwater flow model for the study area. Table 4. 1 evaluates the performance of the ANN models using two criteria: average absolute relative error (AARE) and root mean square error (RMSE). Both criteria show very small errors, confirming the models' strong performance. In summary, the models perform excellently, with minimal errors, showcasing their high accuracy in predicting desired outcomes for the training data. This suggests that the models are reliable for simulating groundwater flow.



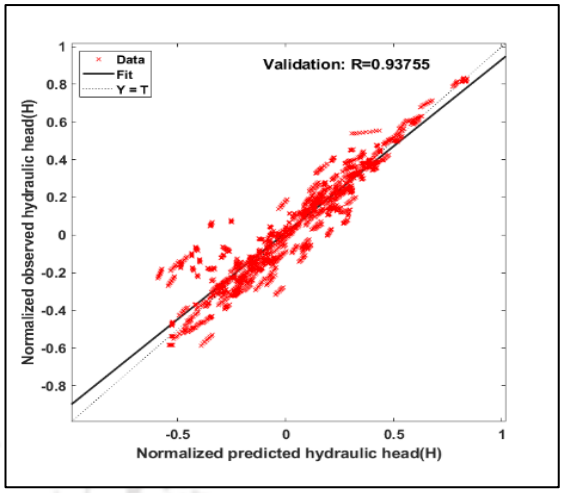
Well 1



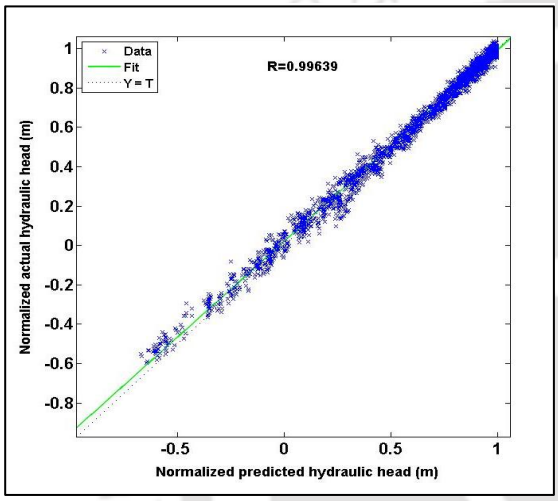
Well 2



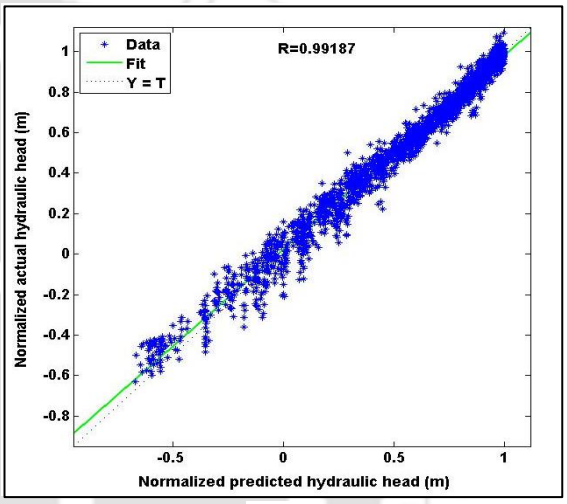
Well 3



Well 4

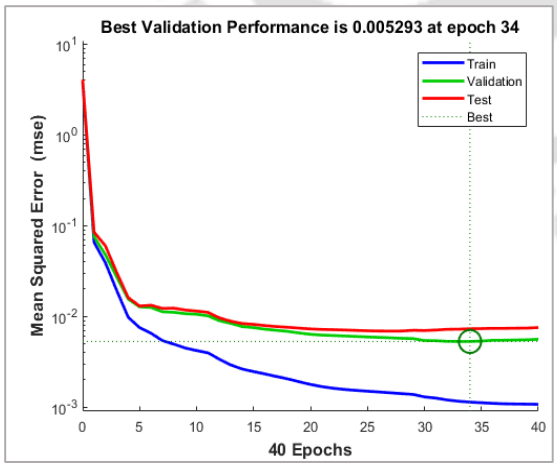


Well 5

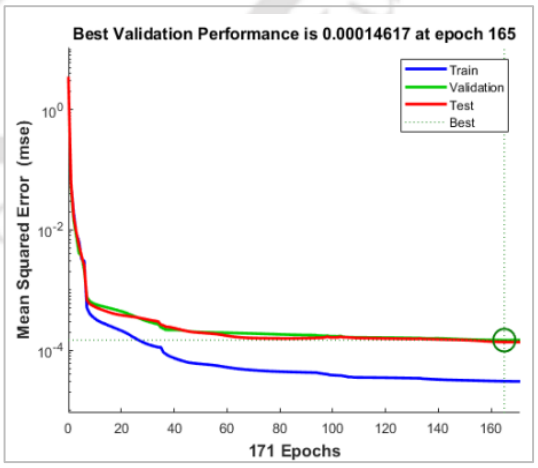


Well6

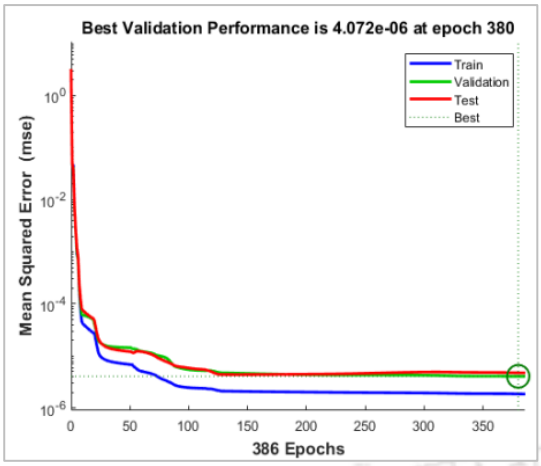
Figure 4. 6: Regression plot of the simulation model with the six ANN models.



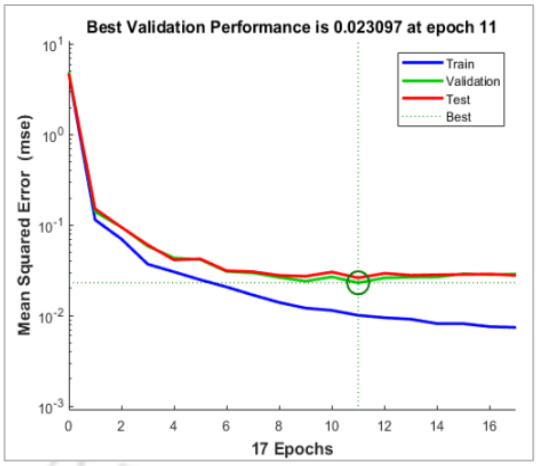
Well 1



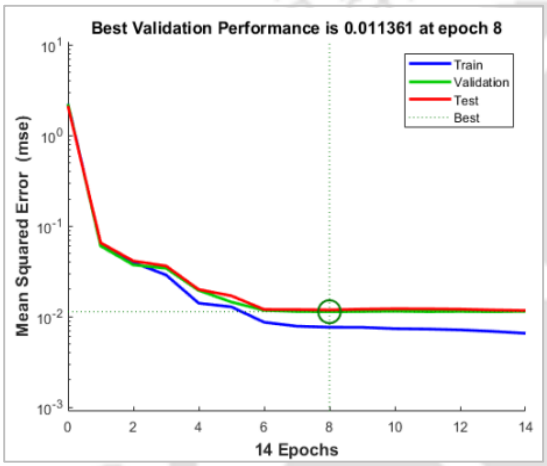
Well 2



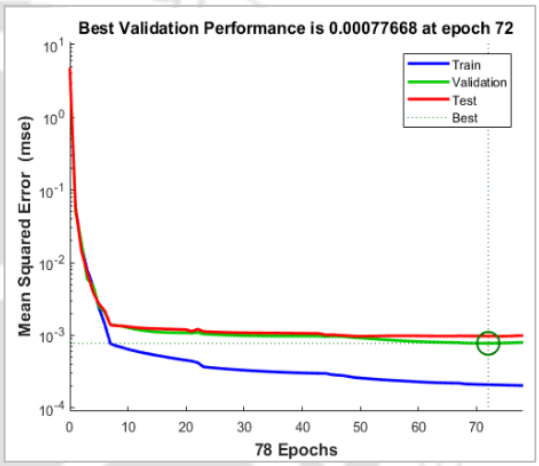
Well 3



Well 4



Well 5



Well 6

Figure 4. 7: Training, testing, and validation plots of the six ANN models.

Table 4. 1: Performance of the trained ANN model based on AARE and RMSE

ANN MODEL	AARE	RMSE
Well 1	0.71	0.33
Well 2	0.68	0.77
Well 3	0.77	0.17
Well 4	0.97	0.34
Well 5	0.73	0.33
Well 6	1.54	0.37

4.3.4.2 Performance of the ANN-SFLA based parameter estimation model:

Shuffled frog leaping algorithm:

The ANN models developed in section 4.3.4.1 are coupled with the optimization model to estimate the flow parameters (i.e., θ_s, θ_r, α, K_s, n, and S_s). Since this study considers a synthetic

problem, the observed hydraulic heads (OH) are taken from the numerical simulation model developed in Chapter 3, and the simulated heads (SH) at different iteration is obtained from the ANN models. The lower and upper limits for these parameters are decided based on the previous experimental evidence and are listed in Table 4.2. The parameters used in SFLA, number of memeplexes is selected as 7, and the number of virtual frogs is selected as the maximum number of variables (6) +1 and equals 7. Therefore, 7 virtual frogs for each memeplex were selected, comprising 49 frogs. The maximum step size was taken as 1 unit. The maximum number of evolutions in each memeplex is 6, the step length coefficient is 2, and the maximum number of iterations is restricted to 200. The predicted flow parameters using the ANN-SFLA-based parameter estimation model are listed in Table 4.2.

Table 4.2: Optimization results 1: ANN-Shuffled frog leaping algorithm

Parameters	$\boldsymbol{\theta_s}$	$\boldsymbol{\theta_r}$	$\boldsymbol{\alpha}$	$\boldsymbol{n}$	$\boldsymbol{K_s}$	$\boldsymbol{S_s}$	**Function value**
Lower limit	0.1	0.009	2	2	0.1	0.0001	6.3084E-05
Upper limit	0.5	0.1	7	15	0.5	0.01	
Actual	0.41	0.055	3.55	4.1	0.678	0.001	
Predicted	0.3992	0.0551	3.4138	4.0812	0.664	0.001001	
Relative Error (%)	2.63	0.182	3.837	0.487	3.44	0.00	

The relative efficiency of the ANN-SFLA model in predicting the flow parameters is also checked by evaluating the relative error with respect to the actual flow parameters. From the model observation, it is observed that the model could predict all the parameters up to a fair degree of accuracy. Considering the relative error among the actual and predicted values, the error ranges from 0.093% to 3.837%. These values are subsequently low and considered within the acceptable range of accuracy. The model converges towards the optimal solution when it reaches an objective function value of 6.3084E-05 after 74 iteration. To compare the model performance, another ANN-Genetic algorithm-based parameter estimation model is developed for estimating the flow parameters.

4.3.4.3 Comparison of SFLA with Genetic Algorithm

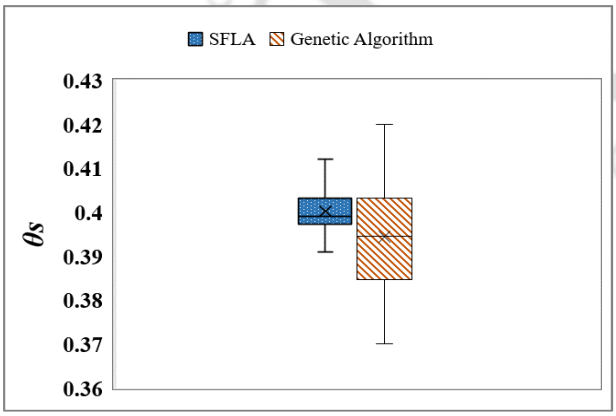
The Genetic algorithm (GA) available in the MATLAB toolbox is used to compare the results with SFLA. Genetic algorithms search for optimal solutions through natural selection and genetic evolution (Abdelgawad, 2009; Cavazzuti, 2013; Mathew, n.d.). In this study, GA uses

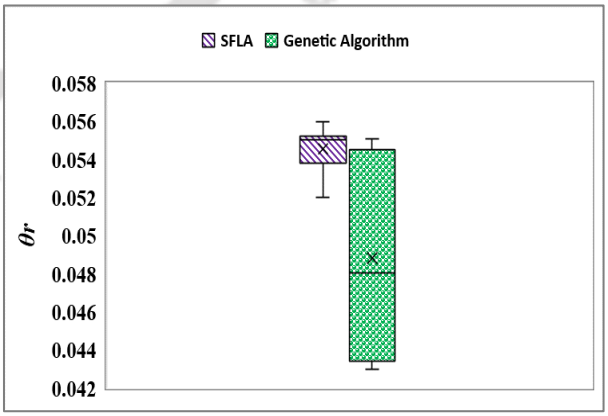
a population size of 50, a maximum generation of 200, a function tolerance of 1×10^{-5}, and a crossover probability of 0.8. Mutation functions are constraint-dependent, and the number of stall generations is 60. The solution obtained using ANN-GA is presented in Table 4. 3

Table 4. 3: Optimization results II: ANN-Genetic algorithm.

Parameters	θ_s	θ_r	α	n	K_s	Ss	Function value
Lower limit	0.1	0.009	2	2	0.1	0.0001	3.31E-03
Upper limit	0.5	0.1	7	15	0.5	0.01	
Actual	0.41	0.055	3.55	4.1	0.678	0.001	
Predicted	0.393	0.043	3.457	4.028	0.672	0.00109	
Relative error (%)	7.31	21.82	2.61	1.756	2.18	9.0	

In this case, GA predicted approximately three parameters, and the remaining three parameters showed near-optimal solutions. For a number of trials considering both the ANN-SFLA and ANN-Genetic algorithms, a number of solutions are generated to verify the accuracy of the proposed algorithms. Figure 4. 8 shows a box plot representing the estimated parameters after 20 trials from both models. The plots show that the average value of α, K_s, and n for both models is very close to the optimal solution. In the ANN-GA model, the estimated value of θ_s, θ_r, and Ss varies with a wide range of values compared to ANN-SFLA. The median value obtained for θ_s is 0.34, θ_r is 0.0135, and Ss is 0.0018 (m^{-1}) using the ANN-GA model, while for the ANN-SFLA model, the median values for θ_s is 0.315, θ_r is 0.011, and Ss=0.0012 (m^{-1}). The values obtained from ANN-SFLA are very close to the optimal solution.





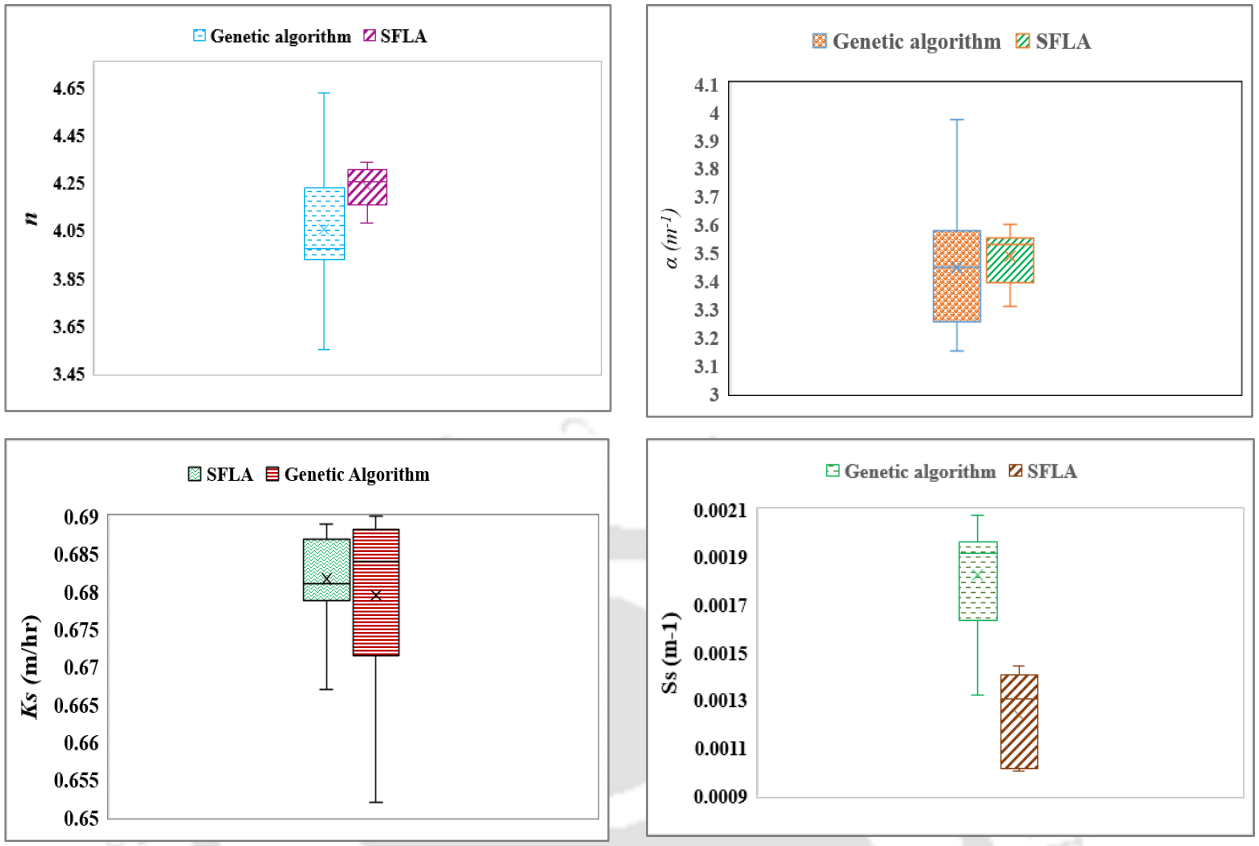


Figure 4. 8: Box plot representing the estimated value of the flow parameters using ANN-Genetic algorithm and ANN-SFLA.

To study the reason for this observation, the variation of the objective function with the iteration for the Genetic Algorithm and Shuffled Frog Leaping Algorithm (SFLA) is plotted in Figure 4.9. A total number of 200 generations is considered for both GA and SFLA. The population size is also relatively similar 50 for GA and 49 for SFLA. Therefore, the total number of function evaluations is almost the same for both algorithms. Figure 4.9 shows that SFLA converges faster and yields better results than GA. While comparing the computational time, it was observed that SFLA based model took 4 hours, 2 minutes and 33 seconds and GA based model took 4 hours, 57 minutes, and 12 seconds. This may be because the memetic evolution is faster and consists of different sets of evolution simultaneously. Genetic evolution consists of the population (a set of solutions) and evolves altogether. On the other hand, memetic evolution follows a different approach where the population is divided into different memeplexes, and each memeplex evolves independently on a population basis. The population is mixed again to communicate so that the global best is updated, and reshuffling is done similarly to continue the evolution into the memeplexes. The problem considered in the study has multiple locally optimal solutions.

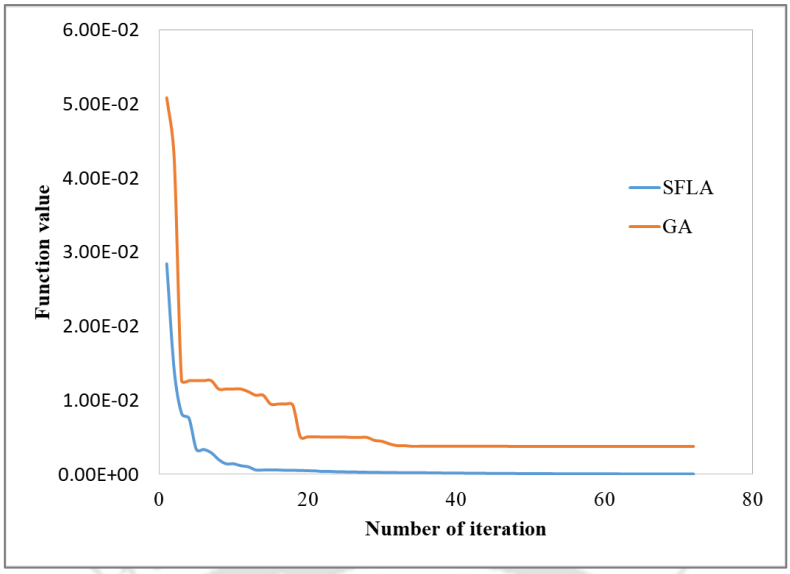


Figure 4.9: Variation of the objective function with iteration for Genetic Algorithm and Shuffled Frog Leaping Algorithm (SFLA).

As such, a large population size must be taken to obtain the optimal global solution. The population size of 50 is considered in GA to compare the results with SFLA. The GA may yield a better solution if we increase the population by more than 50. Thus, we can conclude that due to its faster convergence and better results, SFLA has shown competitive results against GA. As such, ANN-SFLA-based parameter estimation models could be a better alternative to estimate the saturated and unsaturated parameters together.

4.4 Estimation of virus transport parameters using Flower Pollination Algorithm (Model 2):

The second part of this chapter is to develop an effective parameter estimation model to estimate the virus transport parameters in an unconfined aquifer. Previously developed virus transport models available in the literature have typically relied on the constant detachment model to estimate these parameters. However, this approach may not accurately represent real-world conditions where detachment-attachment rates vary based on the saturation level of the medium. The new model proposed in Chapter 3, the "variable detachment model," considers different moisture content in the unsaturated and saturated zones of the aquifer, leading to different adsorption rates. All the models available in the literature are developed either considering equilibrium sorption or kinetic sorption; sometimes, the parameters are overestimated or underestimated in some models. However, an effective parameter estimation model is required to estimate the virus transport parameters considering kinetic sorption. To address this limitation, this study developed a parameter estimation model that considers

variable attachment and detachment rates for two saturation levels. This study aims to estimate the distribution coefficient, inactivation rate, and adsorption rate of viruses at the soil-water and air-water interfaces. Since parameter estimation is a nonconvex problem, the initial attempt to use Genetic Algorithm (GA) did not produce satisfactory results. Therefore, Flower Pollination Algorithms (FPA) are employed in this model as an alternative meta-heuristic optimization algorithm.

4.4.1 Methodology:

In this study, the simulation model developed for virus transport is termed as 'variable detachment model'; this is because the model can simulate the virus concentration at two saturation levels (unsaturated and saturated flow conditions). Since the aquifer's moisture content highly influences the sorption and inactivation mechanism, two different adsorption rates are considered for the unsaturated and saturated zones. Therefore, the attachment and detachment rate of viruses at soil-water interfaces (k_{att}^{s}, k_{det}^{s}), the inactivation rates in a liquid phase, SWI and AWI (μ_w, μ_s, and μ_a), and the distribution coefficient at air-water interfaces (K_D^a) need to be estimated.

4.4.1.1 Problem formulation:

According to Schijven & Hassanizadeh (2000), the ratio of μ_w/μ_s (where μ_w represents the inactivation rate in liquid and μ_s represents the inactivation rate in the solid phase) was found to be close to one. Based on this observation, they assumed that the inactivation rate of the virus in the liquid phase (μ_w) and the solid phase (μ_s) can have the same value. Furthermore, when fitting the value of μ_a (the inactivation rate for unsaturated flow conditions) in corresponding to the breakthrough curves, they noticed a similar trend. Therefore, they also assumed that the value of μ_a could be considered the same as μ_w and μ_s. Since μ_w, μ_s, and μ_a have the same values, one value is estimated instead of three inactivation rates. Therefore, six decision variables ($k_{att}^{s,s}$, $k_{att}^{s,u}$,$k_{det}^{s,s}$, $k_{det}^{s,u}$,K_D^a and μ_w) are considered for estimation. The objective function of this parameter estimation model is to minimize the error function between the simulated and observed virus concentration. Since this problem is a synthetic problem and the field observation data is not available so the observed concentration is taken from the solutions obtained from the virus transport model developed in Chapter 3. And for the simulated concentration, the candidate solutions provided by the optimization model at each iterations is used by the numerical simulation model to generate the simulated concentration. Figure 4. 10 shows the flow chart of the proposed methodology.

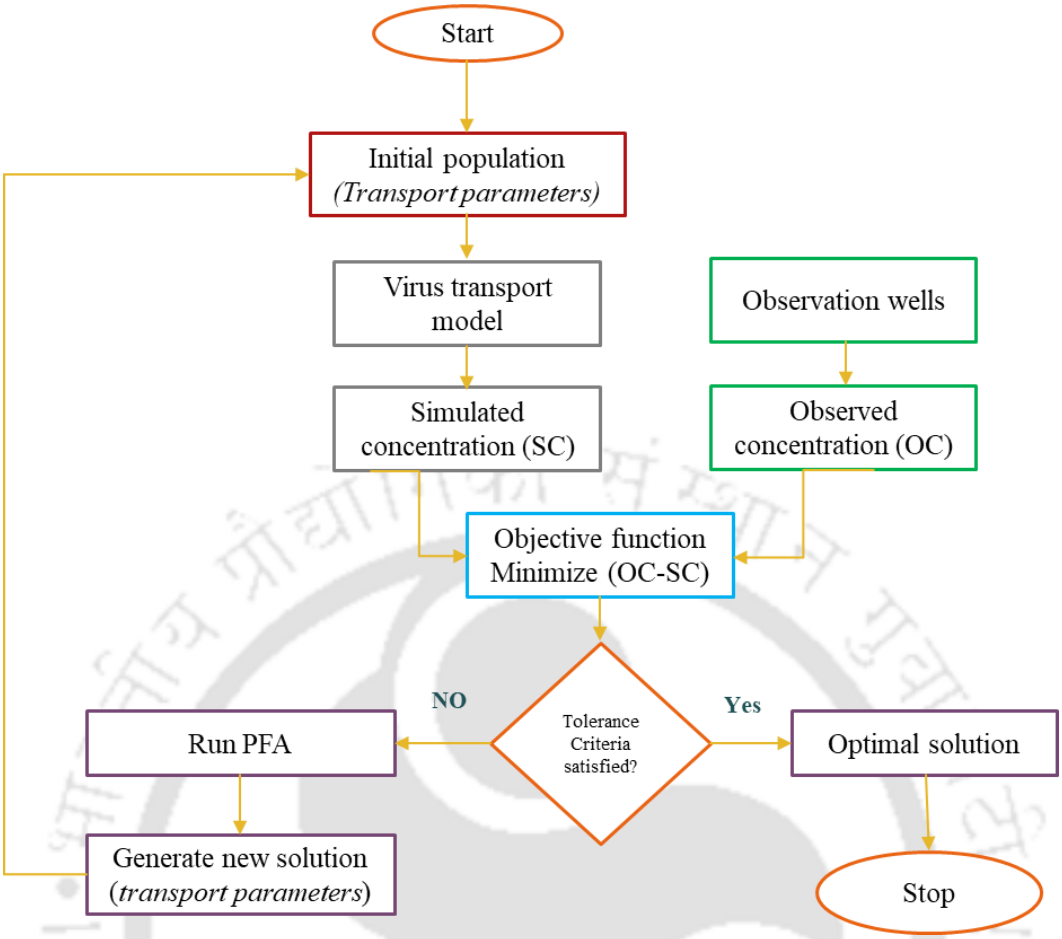


Figure 4. 10:Flowchart of the virus transport parameter estimation model

Minimize: $$f_x = \sum_{j=1}^{M}\sum_{i=1}^{N}\left\|OC_i^j - SC_i^j\right\| \quad (4.6)$$

Subject to:

$$C_W = f\left(k_{att}^{s,s}, k_{\det}^{s,s}, k_{att}^{s,u}, k_{\det}^{s,u}, K_D^a, \mu_w\right)$$

$$k_{att}{}^{s\,min} \le k_{att}{}^{s} \le k_{att}{}^{s\,max}$$

$$k_{det}{}^{s\,min} \le k_{det}{}^{s} \le k_{det}{}^{s\,max}$$

$$K_D{}^{a\,min} \le K_D{}^{a} \le K_D{}^{a\,max}$$

$$\mu_w{}^{min} \le \mu_w \le \mu_w{}^{max}$$

Here, f_x is the objective function for the present optimization model with x number of virus transport parameters; OC_i^j is the observed concentration at i^{th} time step for j^{th} well location; SC_i^j is the simulated concentration at i^{th} time step for j^{th} well location; M is the total number of observation wells and; N is the total number of time steps; $\mathbf{C_W}$ is the concentration vectors

of the simulated concentration; The suffix **'*min*'** and **'*max*'** represents the lower and upper bounds of the virus transport parameters.

4.4.1.2 Virus transport model:

The proposed methodology involves coupling a simulation model of virus transport with an optimization model to estimate the transport parameters. In this case, the simulation model developed in the previous chapter for scenario 1 is chosen as the basis for this parameter estimation model. The simulation model is linked with the optimization model in a manner where the decision variables represent the input virus parameters, and the output is the simulated virus concentration in the aqueous phase at different observation wells. The optimization algorithm will adjust the decision variables iteratively to minimize the discrepancy between the simulated and observed virus concentrations, ultimately finding the best-fit values for the transport parameters. The observation wells are considered near the virus source locations and are shown in Figure 4. 11. The breakthrough curve of the virus concentration obtained at the observation wells from the numerical transport model developed in chapter 3 is represented in Figure 4.12.

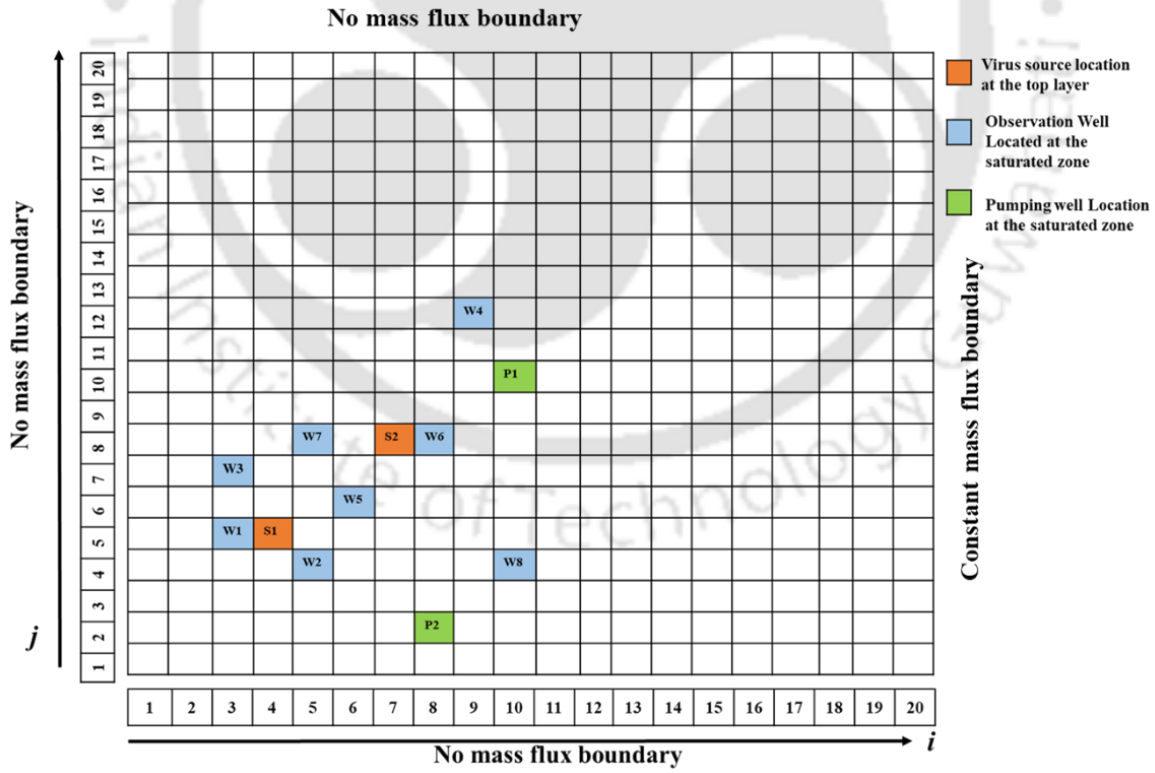


Figure 4. 11: Grid distribution of the virus transport model showing the observation well locations, virus sources, and the pumping well in XY plane.

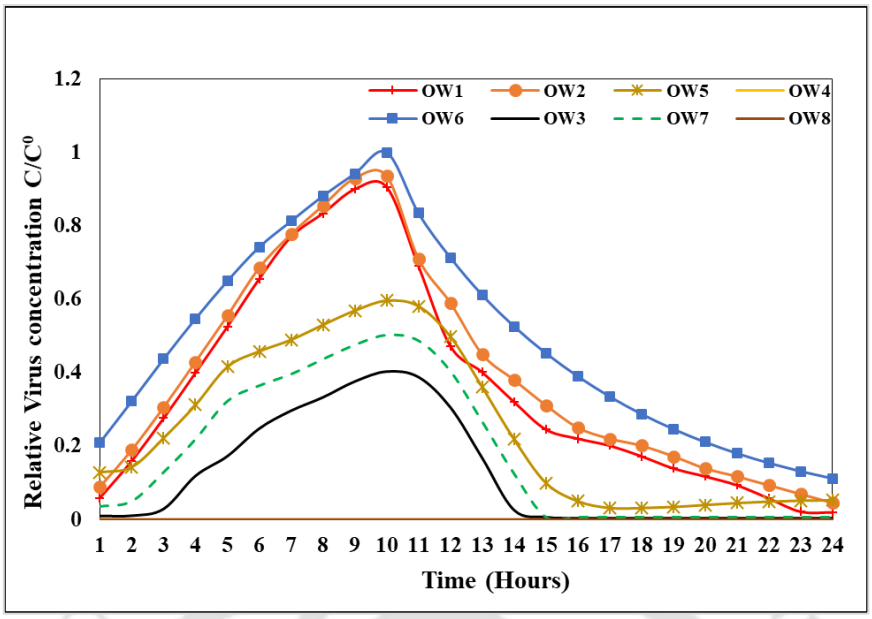


Figure 4.12: Breakthrough curve of the virus concentration for six observation wells.

4.4.1.3 Flower Pollination Algorithm:

The optimization algorithm used to estimate the virus transport parameters is the Flower Pollination Algorithm (FPA). This optimization algorithm takes inspiration from the pollination process in flowering plants. It mimics the two main types of pollination observed in nature: self-pollination and cross-pollination. Self-pollination occurs within short distances and can be seen as a local optimization process. In the context of FPA, this corresponds to the exploration of the search space in the vicinity of the current solution. It involves the optimization of a candidate solution using local search techniques. On the other hand, cross-pollination involves the transfer of pollen between different plants, often facilitated by pollinators like bees, bats, birds, and flies (see Figure 4. 13). This process contributes to global optimization as it allows for genetic diversity and the exploration of distant regions in the search space. In FPA, cross-pollination is simulated through the movement of pollen, which represents candidate solutions, across the search space. The movement of pollen during cross-pollination is modeled using a Lévy flight. A Lévy flight is a random walk following the Lévy distribution, characterized by heavily power-law tails (Shlesinger et al., 1995). This distribution is used to describe anomalous diffusion, where the mean and variance are infinite, resulting in long movements from the current position. By incorporating Lévy flights, FPA enables more efficient exploration of the search space, potentially leading to the discovery of better solutions (Fredriksson, 2010). The steps used in the Flower pollination algorithm is shown in Figure 4. 14 as given by Maseri et al., (2019).

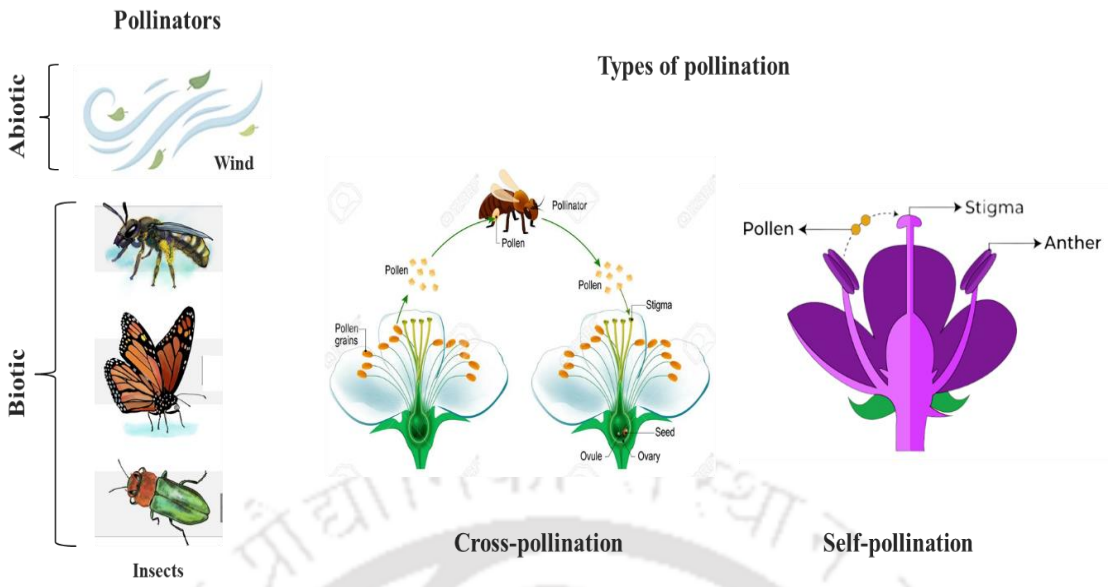


Figure 4. 13: Types of pollinators and pollinations.

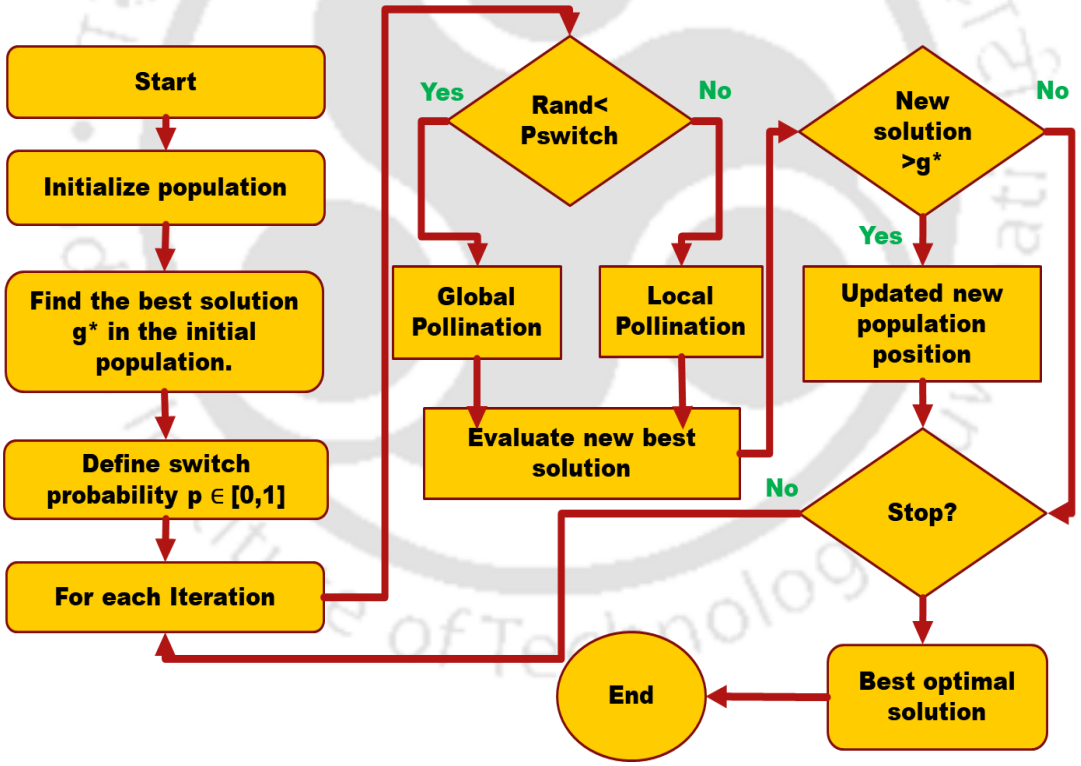


Figure 4. 14: Flowchart of Flower pollination algorithm.

PFA is followed by certain rules as given by (Yang 2012), which are mentioned below:

Rule 1: Global optimization mainly follows the cross-pollination mechanism by considering the biotic pollinators that obey Levy's flight.

Rule 2: Local optimization is caused by Abiotic and self-pollination.

Rule 3: Flower constancy means there will be reproduction probability if two flowers are similar.

Rule 4: Switch probability is considered between ∈ [0, 1], which defines Local or global pollination.

The algorithm starts with randomly generating the initial population, which is evaluated to determine the current best solution. For calculating a new solution, the pollination type should be determined at first according to a predetermined probability p (R4). i.e., a random number r ∈ [0, 1] is generated, and if r is less than p then the global pollination and the flower constancy (R1 and R3) can take place as follows:

$$x_i^{t+1} = x_i^t + L\left(g^t - x_i^t\right) \tag{4.7}$$

Where, x_i^t is a solution *I* at time *t*, g^t is the current best solution, and *L* is a step size drawn from Lévy flight as:

$$L \sim \frac{\lambda\Gamma(\lambda)\sin(\pi\lambda/2)}{\pi} \times \frac{1}{s^{1+\lambda}}, \qquad (S \geq S_0 > 0) \tag{4.8}$$

From equation (4.8), $\Gamma(\lambda)$ represents the standard gamma function, and the distribution is valid for large steps S > 0.

Otherwise, if *r* is greater than *p* then the local pollination and the flower constancy (R2 and R3) are performed as:

$$x_i^{t+1} = x_i^t + \varepsilon\left(x_j^t - x_k^t\right) \tag{4.9}$$

Where x_j^t and x_k^t represent pollen from different flowers of the same species of plant. Thus, mimics the flower constancy in a limited neighborhood. The switch probability or proximity probability is used to switch between common global pollination and intensive local pollination. After that, the current best is updated, and the search iterations resume until the termination criteria. To understand the analysis and steps used in FPA code, I have presented a Pseudo code in Figure 4.15.

1. Initialize the parameters with switch probability $p \in [0\ 1]$;
2. Randomly initialize the population of the flowers;
3. Evaluate the initial population and find the current best solution (g)
4. **while** (termination criteria not satisfied) do
5. **For** each flower
6. **if** rand() < p
7. Global pollination: $x_i^{t+1} = x_i^t + L\left(g^t - x_i^t\right)$
8. **else**
9. Select two random solution x_j^t and x_k^t
10. Local pollination: $x_i^{t+1} = x_i^t + \varepsilon\left(x_j^t - x_k^t\right)$
11. **end if**
12. Evaluate new solutions;
13. Update solutions with better new ones;
14. **end for**
15. Keep the current best solution;
16. **end while**

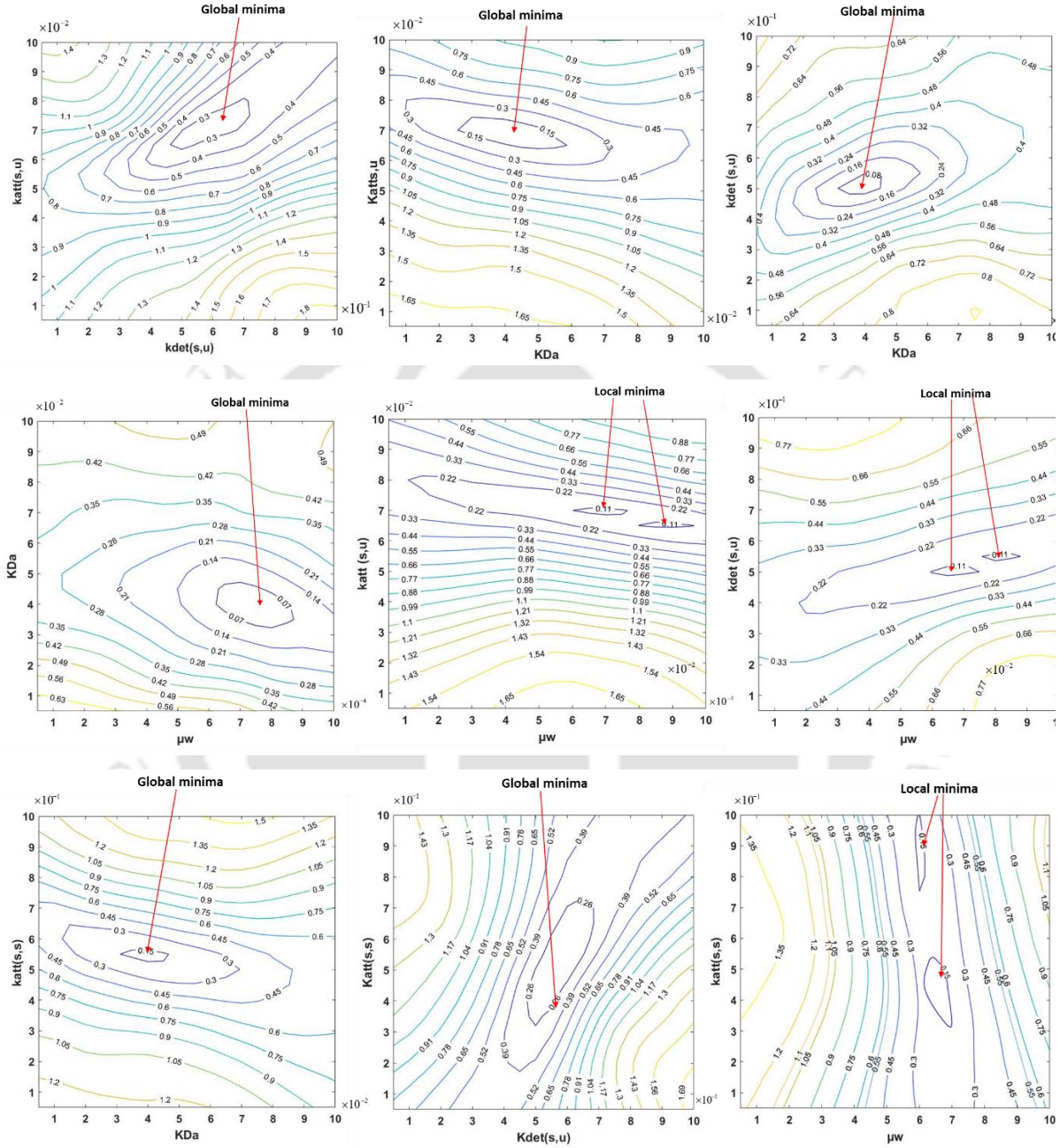
Figure 4.15: Pseudocode for Flower pollination Algorithm (FPA)

4.4.2 Observation from virus transport parameter estimation model:

4.4.2.1 Identification of parametric space:

Non-uniqueness is one of the challenges that can arise in parameter estimation problems. Non-uniqueness occurs when there are multiple sets of parameter values that can produce the same observed data or fit the model equally well. In other words, different combinations of parameter values can result in the same likelihood or goodness-of-fit to the data. Thus, an attempt has been made to study whether the inverse procedure uniquely estimates the unknown parameter sets. In this study, six parameters are selected for estimation; as such, contour plots of the objective function are generated in two-dimensional parametric spaces for twelve different parameter combinations; the result is shown in Figure 4. 16. From the parametric space analysis, it can be observed that when parameter sets such as ($k_{det}^{s,s}$, μ_w), ($k_{att}^{s,s}$, μ_w), ($k_{att}^{s,u}$, μ_w), and ($k_{det}^{s,u}$, μ_w) were considered as unknown, multiple local minima were observed in the parametric space. The other combination of parameter sets shows a single global minimum in the parametric space. Based on these observations, it can be concluded that the presence of adsorption rate and inactivation coefficient in the sorbed phase makes this problem a nonconvex optimization problem. This means that multiple solutions satisfy the given objective function, so the

estimation of the unknown parameter sets becomes non-unique. To overcome the non-uniqueness of the solution, the metaheuristic Flower Pollination Algorithm (FPA) is used as the optimization algorithm in this study.



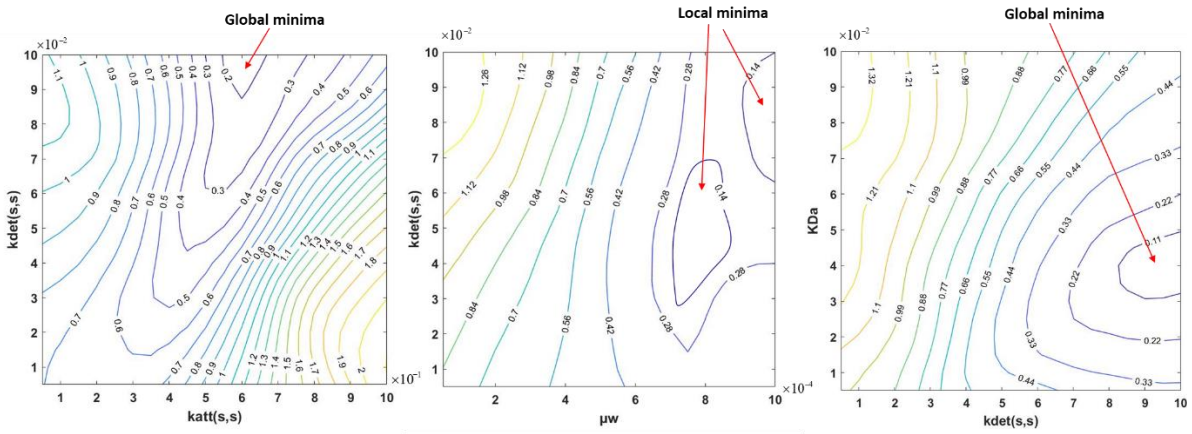


Figure 4. 16: Contour plots of the objective function f in two parameters search space.

4.4.2.2 Performance of the parameter estimation model (Model 2):

The parameters to be estimated for the virus transport process are $k_{att}^{s,s}$, $k_{att}^{s,u}$,$k_{det}^{s,s}$, $k_{det}^{s,u}$, $K_D^{a,}$ and μ_w.

Table 4. 4: Optimization result I: using Genetic Algorithm

Parameters	K_D^a (m)	(100% saturation)		Partially saturated		μw (h^{-1})	Function value
		k_{att}^s(h^{-1})	k_{det}^s (h^{-1})	k_{att}^s(h^{-1})	k_{det}^s(h^{-1})		
Lower limit	0.01	0.1	0.01	0.01	0.1	1E-05	7.35005
Upper limit	0.1	0.9	0.1	0.1	0.9	1E-03	
Actual	0.04	0.594	0.095	0.0695	0.504	7.79 E-04	
Predicted	0.040	0.507	0.088	0.0691	0.499	5.33 E-04	
Relative error (%)	0.0	14.646	7.368	0.575	0.992	31.5789	

Table 4. 5: Optimization results 2: Flower pollination algorithm

Parameters	K_D^a (m)	(100% saturation)		Partially saturated		μw (h^{-1})	Function value
		k_{att}^s (h^{-1})	k_{det}^s (h^{-1})	k_{att}^s (h^{-1})	k_{det}^s (h^{-1})		
Lower limit	0.01	0.1	0.01	0.01	0.1	1E-05	4.095E-04`
Upper limit	0.1	0.9	0.1	0.1	0.9	1E-03	
Actual	0.04	0.594	0.095	0.0695	0.504	7.79E-04	
Predicted	0.041	0.612	0.0949	0.07001	0.51	6.16E-04	
Relative error (%)	2.5	3.030	0.105	0.733	1.190	21.01551	

Observations:

The parametric space study reveals that the parameter estimation model exhibits non-convexity, indicating a complex optimization problem. Initially, Genetic Algorithm (GA) is

employed to address this. The solution obtained from the analysis is given in Table 4. 4. The genetic algorithm does not use gradient information, so it obtains a near-optimal solution rather than the global optimal solution. We have also explored the efficiency of the Flower Pollination Algorithm (FPA) in estimating the virus transport parameters. It has been observed that the solution obtained using FPA provides an actual solution with up to five parameters, but the inactivation rates obtained provide a non-unique solution (ref. Table 4. 5). The function value obtained using FPA is 4.095E-04, which is obtained after 92 iterations.. For a better understanding of the optimal solution, the relative errors of the estimated parameters are evaluated with respect to the actual parameters. The highest relative error among all the estimated parameters is 21.01 %, whereas the lowest is 0.105 %. These relative error values are within acceptable ranges for all the estimated transport parameters except μ_w. As we can see, some solutions may be either overestimated or underestimated. In such cases, it becomes most important to study the sensitivity of each parameter to the model output.

4.5 Comparison of SFLA and FPA:

In this Chapter, the parameter estimation model uses Shuffled Frog Leaping Algorithm (SFLA) for flow parameter and the Flower Pollination Algorithm (FPA) for transport parameter estimation. Determining which algorithm is best for solving a parameter estimation problem between SFLA (Shuffled Frog-Leaping Algorithm) and FPA (Flower Pollination Algorithm) depends on various factors such as the specific characteristics of the problem, the nature of the objective function, and the computational resources available. Both SFLA and FPA are population-based metaheuristic optimization algorithms inspired by natural phenomena but both have distinct mechanisms. SFLA employs a combination of local and global searches through the use of memeplexes and a global shuffle operation to enhance exploration. Whereas, FPA Utilizes global search based on the flower pollination process, where solutions are improved iteratively by exchanging information between different flowers. In the case of the convergence speed, it can vary depending on the problem and parameter settings; however, FPA converges slower compared to SFLA (Kok et al., 2022). Which was also observed while estimating the flow and transport parameters in this study.

4.6 SOBOL'S sensitivity analysis

SOBOL's global sensitivity method is used to study the most influential flow and transport parameters in developed three dimensional flows and virus transport models. In this method, the variance decomposition of output, based on an extensive sampling of input parameters, can

be used to determine the effect of each parameter and its interactions. This method is based on the variance decomposition approach and is capable of non-linear and non-monotonic models. The steps followed by SOBOL's global sensitivity analysis are presented in Figure 4.17.

4.6.1 Methodology

The models can be represented in functional form as follows:

$$y = f(x) = f(x_1, x_2, \ldots\ldots\ldots\ldots\ldots, x_p) \tag{4.10}$$

Where y represents the goodness of fit metric of model output, $x = (x_1, x_2, \ldots\ldots, x_p)$ is set of input parameters. Since SOBOL's method is a variance decomposition approach, the total variance of function f is represented by V(y). V(y) is decomposed into different components considering the variance of individual parameters and their interactions.

$$V(y) = \sum_i V_i + \sum_{i<j} V_{ij} + \sum_{i<j<k} V_{ijk} + \ldots\ldots\ldots\ldots\ldots\ldots + V_{12\ldots p} \tag{4.11}$$

Now, the SOBOL's sensitivity indices for different orders are obtained considering the percentage contribution to the total variance V.

The first order sensitivity indices (S_i) on y are then defined as:

$$S_i = \frac{V_i}{V} \tag{4.12}$$

The second order indices (S_{ij}) on y due to the direct effect between the two parameters x_i and x_j is given by:

$$S_{ij} = \frac{V_{ij}}{V} \tag{4.13}$$

The total-order indices (S_t) account for the direct effects between one parameter x_i with the other parameters and are given by equation (4.14):

$$S_t = 1 - \frac{V_{\sim i}}{V} \tag{4.14}$$

Where V_i represents the variance due to i^{th} parameter, V_{ij} represents the variance between two parameters (x_i and x_j). $V_{\sim I}$ represents the total variance due to all parameters except the one parameter for which we are calculating the total order indices. The Monte Carlo approximation is used to find the variance in equation (4.11) by using the following equation

$$\hat{f}_0 = \frac{1}{n}\sum_{S=1}^{n} f(x_s) \tag{4.15}$$

$$\hat{V} = \frac{1}{n}\sum_{S=1}^{n} f^2(x_s) - \hat{f}_0^2 \tag{4.16}$$

$$\hat{V}V_i = \frac{1}{n}\sum_{S=1}^{n} f\left(x_s^{(a)}\right) f\left(x_{(\sim i)s}^{(b)} - x_{is}^{(a)}\right) - \hat{f}_0^2 \tag{4.17}$$

$$\hat{V}V_{ij}^{c} = \frac{1}{n}\sum_{S=1}^{n} f\left(x_s^{(a)}\right) f\left(x_{(\sim i,\sim j)s}^{(b)} - x_{(i,j)s}^{(a)}\right) - \hat{f}_0^2 \tag{4.18}$$

$$\hat{V}V_{ij} = \hat{V}V_{ij}^{c} - \hat{V}V_i - \hat{V}V_j \tag{4.19}$$

$$\hat{V}_{\sim i} = \frac{1}{n}\sum_{S=1}^{n} f\left(x_s^{(a)}\right) f\left(x_{(\sim i)s}^{(a)}, x_{is}^{(b)}\right) - \hat{f}_0^2 \tag{4.20}$$

Where n represents sample size, x_s is the sampled individual in the scaled unit hypercube, and superscripts (a) and (b) represent two different samples. All of the parameters take their values from sample (a) and are represented by $x_s^{(a)}$. The variables $x_{is}^{(a)}$ and $x_{is}^{(b)}$ denote that parameter x_{is} uses the sampled values in a sample (a) and (b), respectively. The symbols $x_{\sim is}^{(a)}$ and $x_{\sim is}^{(b)}$ represent cases when all of the parameters except for x_{is} use the sampled values in sample (a) and (b), respectively. The symbol $x_{(ij)s}^{(a)}$ represents parameters x_{is} and x_{js} with sampled values in sample (a). Finally, $x_{(\sim i\sim j)s}^{(a)}$ represents the case when all of the parameters except for x_{is} and x_{js} utilize sampled values from sample (b).

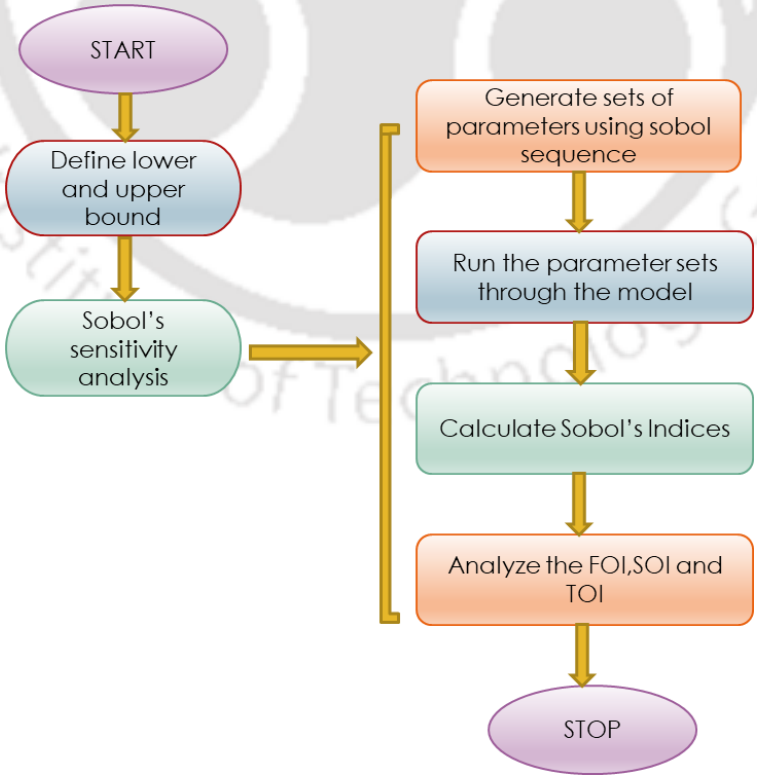
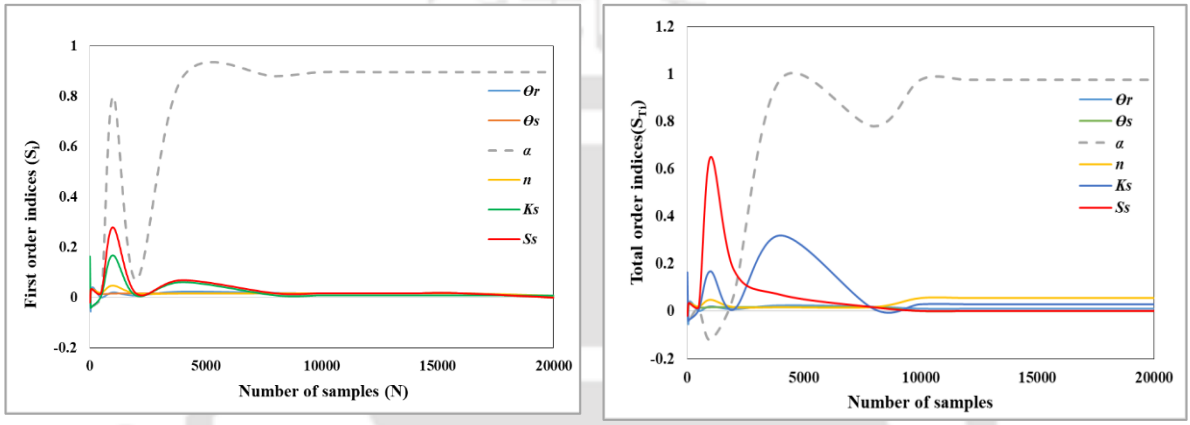


Figure 4.17:Steps of Sobol's sensitivity.

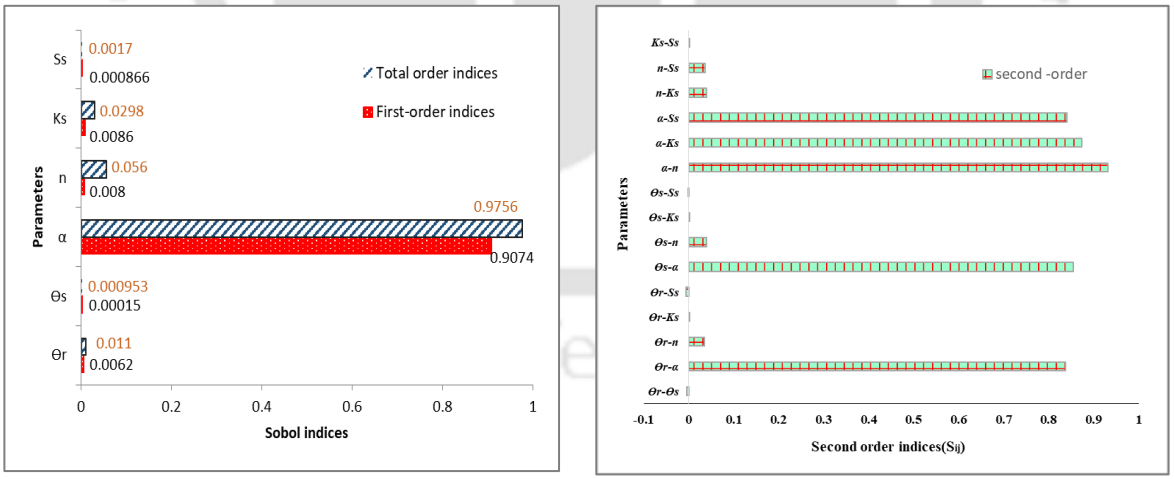
4.6.2 Flow parameters

Selection of the sample size is one of the most significant steps while carrying out SOBOL's sensitivity analysis. To start the analysis, the SOBOL indices (total order indices and first-order indices) are calculated with flow parameters as input and hydraulic head as output with different sample sizes. The most suitable sample sizes are then selected as shown in Figure 4.18. As the sample size increases beyond 10000, the values of SOBOL's indices do not change. This means that at least 10000 samples should be considered while performing the sensitivity analysis in this study.



(a) (b)

Figure 4.18: Evolution of the (a) first-order (*Si*) and (b) total-order (*St*) sensitivity of the flow parameters obtained for different sample size.



(a) (b)

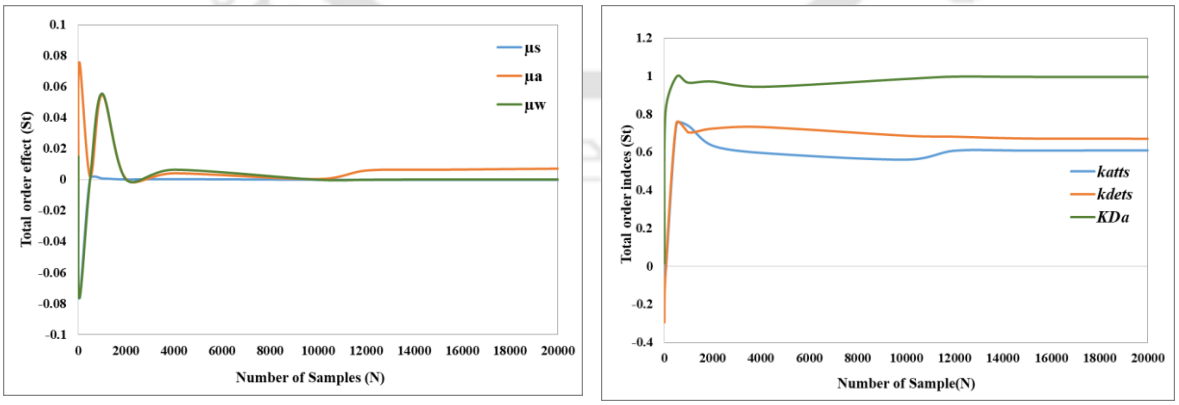
Figure 4.19: Sensitivity analysis of (a) first-order indices and total indices and (b) second-order indices for groundwater flow model.

Using Equations (4.12) to (4.14) the effect of all the flow parameters on the model output using Sobol's First order indices (S_i), Second order indices (S_{ij}), and Total order indices (S_t) are

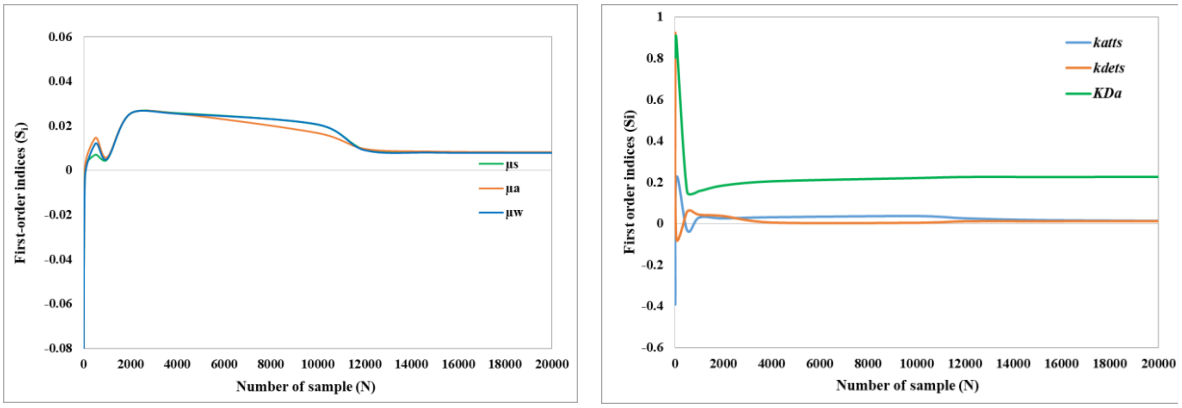
calculated and are shown in Figure 4.19. The analysis is carried out considering a sample size of 15000.When the value of the S_i and S_t approaches to 1, this means that the parameter is highly sensitive. On the other hand, the value of S_i should always be less than S_t. In this study, it was observed that a high value of S_i and S_t (>0.9) is observed for Van Genuchten's fitting parameter (α), which means that the parameter (α) is a highly sensitive input parameter. The second-order Sobol indices are also determined to understand the influence of two parameter to the model output. In this flow domain, the highest value of S_{ij} is obtained for (α-n), followed by (α- θ_s),(α- θ_r), (α-Ss) and (α-Ks).This result indicates that α is the most sensitive input parameter, which shows the highest value when interacting with the other parameters. These findings indicated that the hydraulic head obtained from the model output had a synchronized effect when the parameter **α** interacted with the other flow parameters, which was not possible to observe during the FOI calculation.

4.6.3 Virus transport parameters

A similar process has been carried out to understand the sensitivity of the transport parameters as it is done for the flow parameters. Initially, the best suitable sample size for this sensitivity analysis is selected by calculating the sensitivity indices using all the transport parameters as input and the virus concentration as the output for different sample sizes. In this study, six parameters were selected for sensitivity analysis. They are inactivation rate (μ_w), attachment and detachment rate co-efficient in soil-water interfaces for both saturated and unsaturated zones ($K_{det}^{s,u}$,$K_{att}^{s,u,}$, $k_{det}^{s,s}$ and $k_{att}^{s,s}$), and distribution coefficient at air-water interfaces (K_D^a). The result obtained for all the transport parameters with different sample sizes is shown in Figure 4.20.

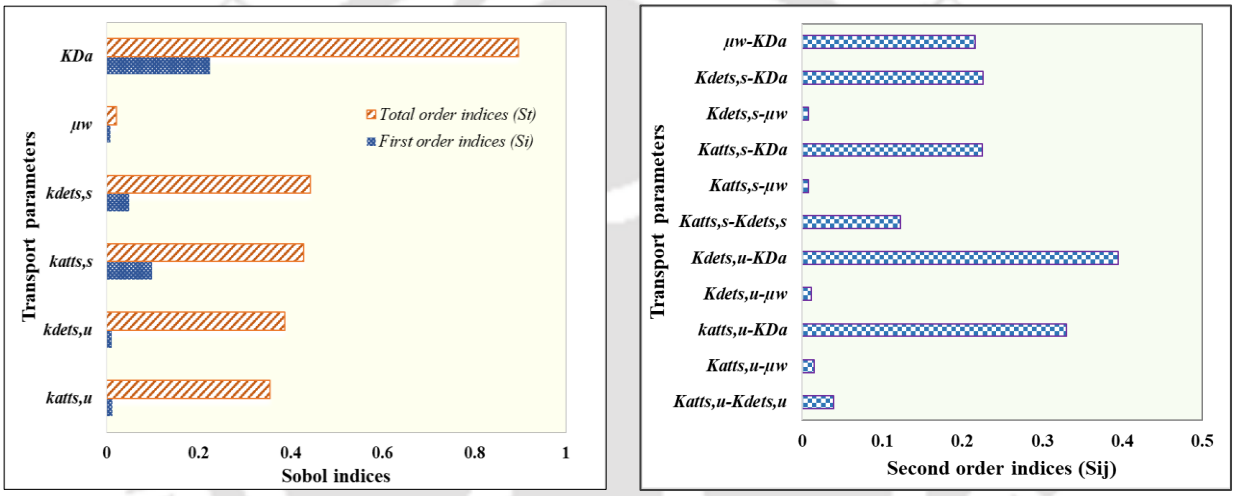


(a)



(b)

Figure 4.20:Evolution of the (a) first-order (Si) and (b) total-order (St) sensitivity of the transport parameters obtained for different sample size



(a) (b)

Figure 4.21: Sensitivity analysis of (a) first-order indices and total indices and (b) second-order indices for virus transport parameters.

In this case, the analysis reveals that when the sample size reaches 14,000, the values of SOBOL's indices no longer change appreciably with further increases in sample size. This suggests that a sample size of at least 14,000 is necessary to conduct the sensitivity analysis reliably. In this study, a sample size of 18000 is selected, and SOBOL's First order indices (S_i), Second order indices (S_{ij}), and Total order indices (S_t) are calculated using equation (4.12) to (4.14). Figure 4.21 shows that a high value of S_i (>0.2) and S_t (>0.9) is observed for distribution coefficient at air-water interfaces (K_D^a) followed by attachment and detachment co-efficient in soil-water interfaces ($k_{det}^{s,s}$ and $k_{att}^{s,s}$), which means that the parameter (K_D^a) is a highly sensitive input parameter. The second-order indices provide the effect of the interactions among the two input parameters on the model output. The second-order SOBOL's indices are also

determined for all the transport parameters to understand the interaction of each parameter with the others (ref. Figure 4.21). The result shows that the highest value of S_{ij} is obtained for the combination of all parameters with K_D^a, among which the highest correlation is seen in K_D^a-$k_{det}^{s,u}$ (>0.39) followed by K_D^a-$k_{att}^{s,u}$ (>0.33). In this study domain, the inactivation rate (μ_w) is found to be less sensitive than the other parameters. Based on the comprehensive investigation of sensitivity analysis in this study area, the distribution coefficient at air-water interfaces (K_D^a) emerges as the significant input parameter. This implies that K_D^a plays a crucial role in determining the behavior and fate of viruses in the unsaturated zone. Also, the attachment ($k_{att}^{s,u}$) and detachment ($k_{att}^{s,u}$) rate coefficient in soil-water interfaces show relative importance to the model output. Since the inactivation rate (μ_w) shows less sensitivity in this study domain, the estimated value using the parameter estimation model is feasible, as a small change in this parameter will not affect the model output.

4.7 Conclusion

This chapter proposes an effective methodology for estimating unsaturated and saturated flow parameters together in a single inverse optimization model. To achieve computational efficiency and accurate predictions, this parameter estimation model utilizes the Shuffled Frog Leaping Algorithm (SFLA) in combination with Artificial Neural Network (ANN). The conclusion drawn from this model are:

- The analysis shows that SFLA enables faster convergence and produces better results than the Genetic Algorithm (GA) with exhibiting minimal relative errors.
- In contrast, the ANN-SFLA-based parameter estimation model is a preferable approach for solving this type of parameter estimation problem.
- Additionally, a sensitivity analysis reveals that the fitting parameter **α** is highly sensitive in the developed groundwater flow model, particularly when it interacts with other parameters.

The second part of this chapter presented the virus transport parameters estimation model in a variable detachment model.

- Initially, an observation was made where the parametric search space analysis revealed numerous local optimal solutions, leading to non-unique optimal solutions when estimating all parameters simultaneously in the model.

- Genetic Algorithm (GA) is applied and it mostly converged to near-global optimal solutions. Later Flower Pollination Algorithm (FPA) was employed and it indicated that the solutions for five transport parameters were accurate, but the inactivation rates obtained were near-optimal.
- Thus, Flower pollination algorithm, on the other hand, demonstrates effectiveness in estimating virus transport parameters, though the model has multiple non-unique optimal solutions.
- While analyzing the global sensitivity, the inactivation rates (μ_a,μ_s,μ_w) show less sensitive input parameters for developing the virus transport model in unsaturated - saturated zone.

CHAPTER 5

IDENTIFICATION OF UNKNOWN VIRUS SOURCES IN THE UNSATURATED AND SATURATED ZONE

5.1 Overview

A methodology is described in this chapter for determining the virus sources iteratively. The problems and methods for identifying groundwater pollution sources are briefly discussed, considering the source location, input source fluxes, and strength as unknown. This chapter presents two models, one identifying the source location in two-dimensional space and another in three-dimensional space considering sources in both unsaturated and saturated zones. Finally, it concludes with a model performance of the source identification showing the effect of measurement error in the observed data.

5.2 Introduction

Identifying the pollution source location and the magnitude of the pollution fluxes poses a significant challenge to the hydrogeologist. A literature review shows that different simulation-optimization techniques have been developed for identifying the sources of groundwater pollution. Due to the limitations of the repose matrix and embedded approaches, linked simulation optimization is a more effective method for managing large aquifer areas. This study uses the linked simulation-optimization model to identify the virus sources in the unconfined aquifer by minimizing the error between the simulated and observed concentrations. The concentrations measured at different observation wells in the field are the observed virus concentrations. Since this study considers a synthetic model, the observed concentration is obtained using the simulation model, keeping all the source locations and concentration flux known. Later, this model is linked with the optimization model to get the simulated concentration at each iteration, keeping the source location and fluxes as decision variables. Using this approach, the optimization model repeatedly calls the Groundwater Transport simulation model to provide the simulated concentration. This study uses the Shuffled frog leaping algorithm (SFLA) in the source identification model as an optimization algorithm. The source identification model available in the literature has two shortcomings: (a) the model could identify the virus sources in a saturated zone, and (b) it can identify the source location in a two-dimensional search space. To overcome such limitation, this study provides two models where (a) the model can identify the source locations in an unsaturated zone in a two-

dimensional search space and (b) identify the source location in both saturated and unsaturated zones in three-dimensional space.

5.3 Methodology

The proposed methodology provides a linked simulation-optimization model, where the objective function is used to minimize the error between simulated and observed virus concentrations. Since the unsaturated zone is present in the study area, the coupled flow and virus transport model is linked with the optimization model. This is because the area is mainly dry in the unsaturated zone, and the contaminant will only move along with the flow velocity. Thus, keeping this in mind, this model considers three decision variables- (a) input flow flux (*If*), (b) virus source locations (*Sl*), and (c) source strength (*Sf*) for each source. So, the total number of variables is based on the '*n*' number of sources present in the study area; that is, the total number of decision variables equals to $n \times 3$. The model developed by Rajeev Gandhi et al., (2017) can be used to obtain the number of sources. The study put forward the number of decision variables and develop the source identification model using the methodology shown in Figure 5.1. Initially, the optimization model generates a set of solutions; from the solution, the input flow flux (*If*) and source location (*Sl*) are used to develop the flow model. This flow model will provide the moisture content and velocity flux at different locations and times. The solution obtained from the flow model is incorporated into the virus transport model. Later, the source locations (*Sl*) and source strength (*Sf)* provided by the optimization model are considered as input to run the virus transport model. The virus transport model will generate the simulated virus concentration at the observation well location based on the solution set provided by the optimization model. The objective function is calculated using the simulated concentration and the observed concentration. If the termination criteria are satisfied, the iteration will be terminated, and the optimal solution will be displayed. If not, the optimization model will again provide another set of solutions to the numerical simulation model to generate the simulated concentration. This process is continued until the termination criteria are satisfied.

The objective function for this optimization model is given by

$$\text{Minimize } f_x = \sum_{j=1}^{M}\sum_{i=1}^{N}\left|OC_{o,i}^{j} - SC_{s,i}^{j}\right| \tag{5.1}$$

$$\text{Subject to } C_W = f\left(If, Sl, Sf\right) \tag{5.2}$$

$$If_{\min} \le If \le If_{\max} \tag{5.3}$$

$$Sl_{\min} \le Sl \le Sl_{\max} \tag{5.4}$$

$$Sf_{\min} \le Sf \le Sf_{\max} \tag{5.5}$$

Here, f_x is the objective function for the present optimization model with x number of virus sources; $OC_{o,i}^{j}$ is the observed concentration at i^{th} time step for j^{th} well location; $SC_{s,i}^{j}$ is the simulated concentration at i^{th} time step for j^{th} well location; M is the total number of observation wells and; N is the total number of time steps; C_W is the concentration vectors of the simulated concentration; If represents the inflow flux applied in the groundwater flow model such that $If = [If_1, If_2 \ldots If_x]^T$; Sf represents the virus source fluxes such that $Sf = [Sf_1, Sf_2 \ldots Sf_x]^T$; Sl represents the virus source location vectors such that $Sl=[Sl_1, Sl_2 \ldots Sl_x]^T$; The suffix '$min$' and '$max$' represent the lower and upper bounds of the source flux, source location, and inflow flux in the source identification model.

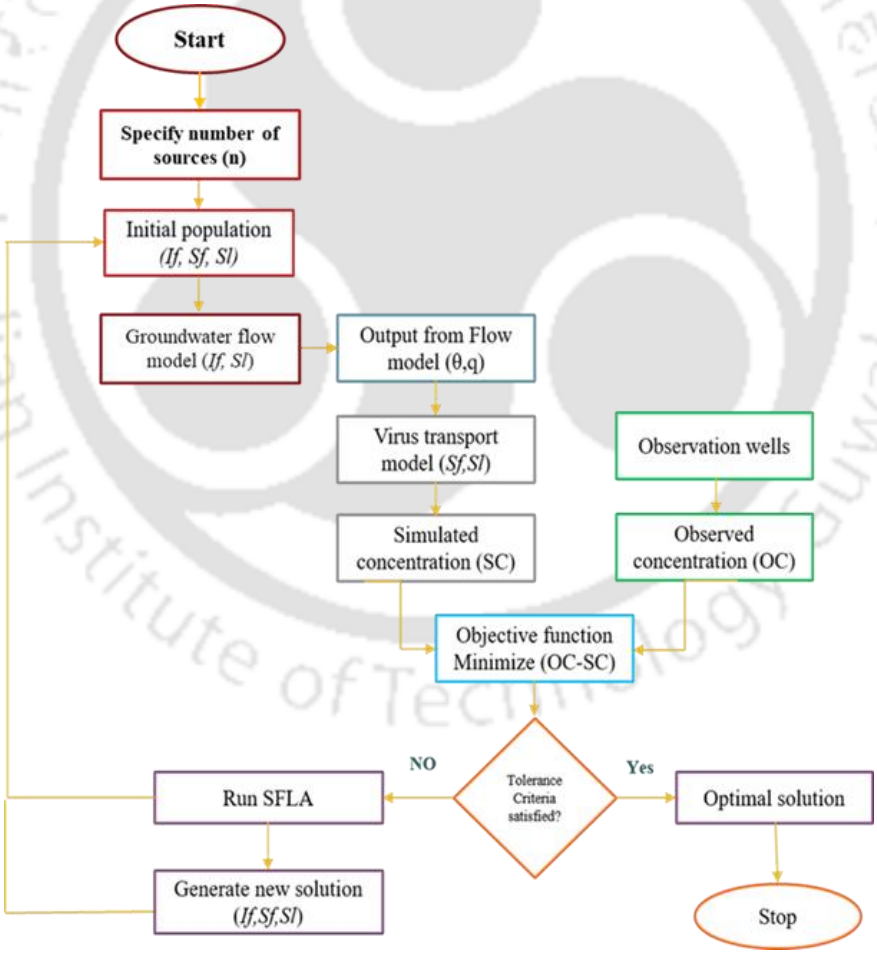


Figure 5.1: Flowchart showing the source identification model

5.3.1 Measurement error of observed data

Measuring virus concentration in the field can be challenging, and accuracy may be affected by several factors, such as the quality of the sample, the sensitivity of the testing method, and the presence of interfering substances. Sampling techniques may not be standardized, and there may be variability in how samples are collected and prepared, which can affect the accuracy of the measurements. The proposed methodology mentions that the observed concentration must be considered to develop any source identification model. As a result, a study has been conducted to test how measurement error affects virus source identification. Different error levels have been introduced to the observed concentrations to assess the impact of measurement error. For the measurement error, the following equation is used (Mahar and Datta, 2001).

$$PC_{o,i}^{j} = C_{0,i}^{j}(1+e) \tag{5.6}$$

Where $PC_{o,i}^{j}$ represents perturbed simulated concentration value, $C_{0,i}^{j}$ represents the observed concentration value, and ***e*** represents the error term. To calculate the error term (e), we use the normal distribution with sigma (σ) and zero mean (μ= 0). Error magnitudes range from 0.05 to 0.2 (R. M. Singh & Datta, 2006). According to them, $\sigma \le 0.1$, it denotes a low noise level, $0.1 \le \sigma \le 0.15$ signifies moderate noise levels, and $\sigma \ge 0.15$ indicates high noise levels. When contaminated concentrations are measured without adding the ***e*** term, the measurements have no noise. Assuming no measurement error exists in the observed virus concentration, the source identification model provides the most accurate solution. Nevertheless, it is different from the realistic case, as explained above. We recommended incorporating the error term in the observed concentration to check the model's effectiveness in actual circumstances. Based on the proposed methodology, the observed concentrations are adjusted with different noise levels to identify virus sources and source fluxes.

5.3.2 Performance evaluation criteria

Relative error (RE) is used to evaluate the performance of the proposed optimization-simulation model (Borah & Bhattacharjya, 2015). Following is the formula for calculating the relative error (RE) criterion between the estimated and actual source fluxes source strength:

$$RE = \frac{\left|Ef_{I,v} - Af_{I,v}\right|}{Af_{I,v}} \times 100 \tag{5.7}$$

Where $Ef_{I,v}$ and $Af_{I,v}$ are the estimated and the actual fluxes for virus sources (v) and inflow fluxes (I), respectively.

5.4 Performance of the source identification model

This study uses the example problems developed in Chapter 3 to evaluate the performance of the proposed source identification model. The unknown variables are input source flux (*If*), Source locations (*Sl*), and source strength (*Sf*).

5.4.1 Virus source identification in two-dimensional space considering source location in unsaturated zone (Model 1):

As developed in Chapter 3, the first scenario is developed by considering two sources due to leakage from sewer lines, and the sources are present in the unsaturated zone. This model is developed by assuming a number of sources (n) = 2, and the number of decision variables is $(n \times 3 = 6)$. It also assumes that it is present at the top surface of the unsaturated zone that is at the first layer so that the model can search the location in two-dimensional space. Here, 16 observation wells are considered to collect the observed concentration from the groundwater (that is, from the saturated zone). Figure 5. 2 shows the arrangement of the source location, observation wells, and pumping well location considered in scenario 1.

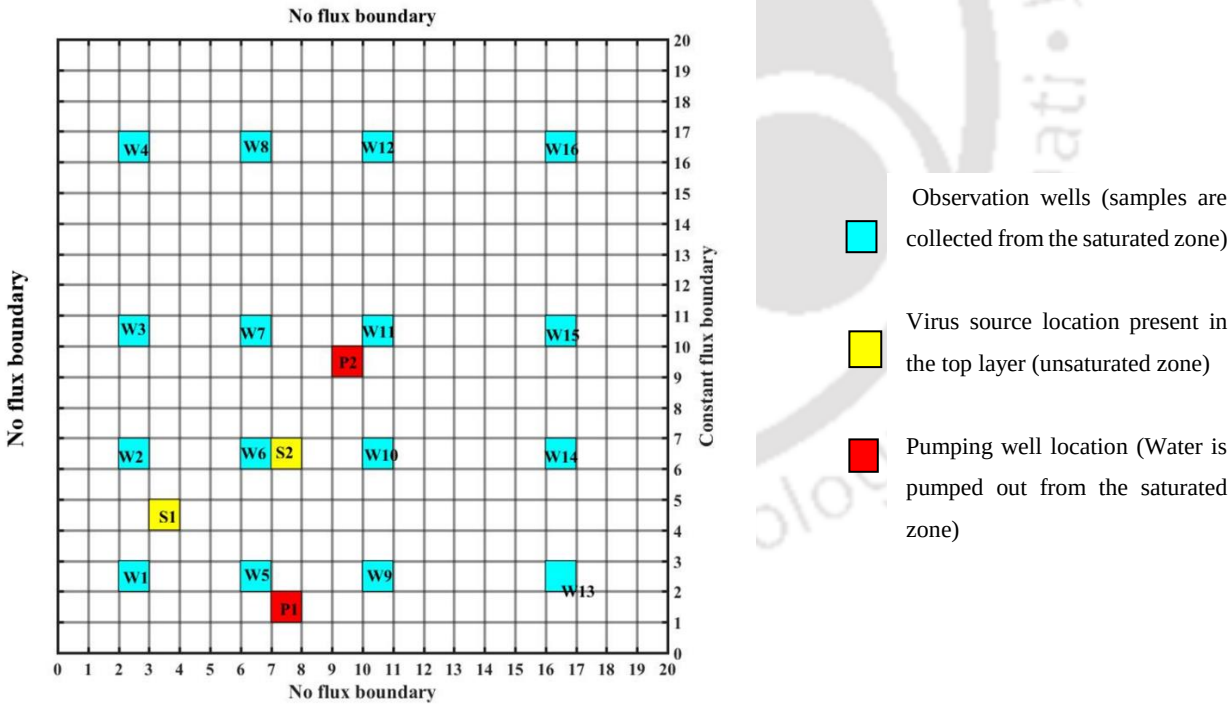


Figure 5. 2: Study area showing the actual source locations, observation well locations and pumping wells of scenario 1.

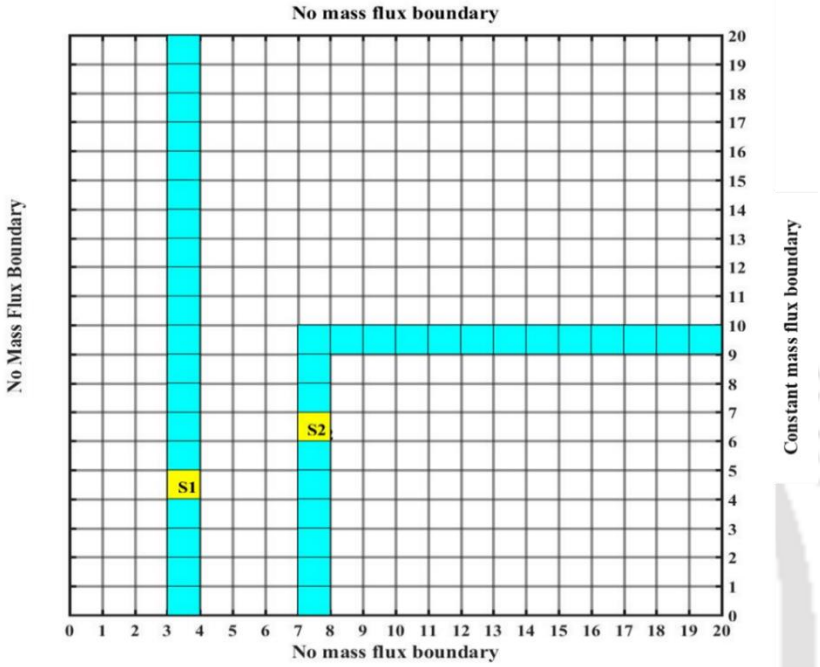
5.4.1.1 Conversion of 2D domain to 1-D domain:

To minimize the variable complexity within a two-dimensional search space, we convert the XY plane coordinates into a singular dimension by assigning grid IDs to each grid, effectively pairing two coordinates with a single source location. By doing so, one can identify the grid

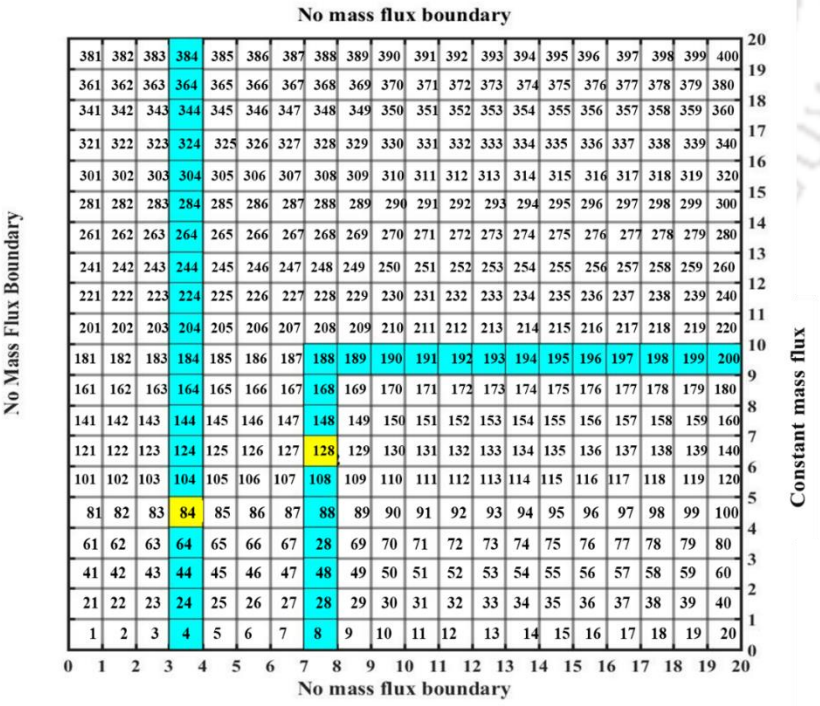
IDs instead of identifying the coordinates corresponding to the *'n'* number of sources. The XY plan can be assigned to grid IDs for two-dimensional search space using the equation (5.8).

$$IDs = (y-1)Ny + x \tag{5.8}$$

Here, *Nx* and *Ny* are the numbers of grid sizes along the X-axis and Y-axis, and *x* and *y* are the coordinates along the x-y plane.



(a) 2-D search space showing the source location.



(b) 2D-search space converted into 1D search space after providing grid IDs.

Figure 5. 3: Grid distribution at the top layer in X-Y plane for Model 1

For Model 1, we assumed that the z-coordinate is known that is Z=1, so the source location in XY plan needs to be identified. Now, these two source locations are marked with grid IDs by using the equation (5.8). In this case, N_x= 20 and N_y=20, so the total number of grids is 400, as shown in Figure 5.3. So, the model must detect the virus source location in between $1 \leq Sl \leq 400$. In this study area, the source locations are mainly present at coordinates Sl_1(4,5,1) and Sl_2(8,7,1), and the corresponding grid IDs are Sl_1= 84 and Sl_2= 128.

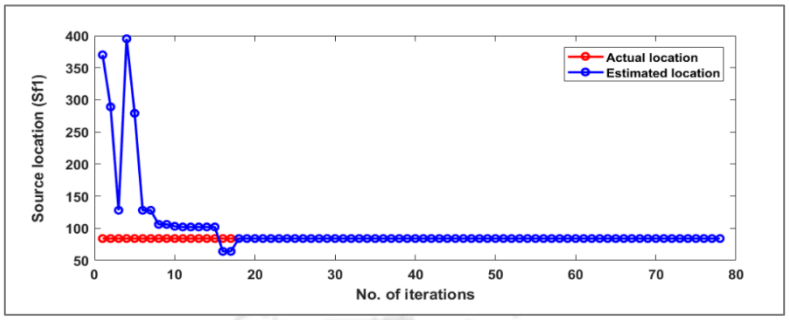
5.4.1.2 Shuffled frog leaping algorithm:

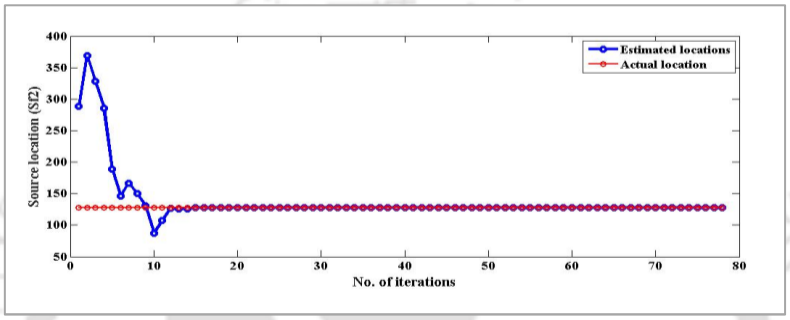
The coupled flow and virus transport model is linked with the optimization model to develop the source identification model. Due to the non-linear nature of the response function, gradient-based classical optimization algorithms are sometimes incapable of finding the global optimal solution to the problem. Furthermore, meta-heuristic optimization algorithms have proven effective at solving such inverse optimization models. This study uses a meta-heuristic optimization algorithm, the Shuffled frog leaping algorithm (SFLA), to identify the virus sources. The steps involved in SFLA are explained in detail in Chapter 4. As per the algorithm, the parameters that need to be defined are the number of memeplexes selected as 7, and the number of virtual frogs chosen as the maximum number of variables plus one (6+1) which equals to 7. Therefore, 7 virtual frogs for each memeplex were selected, comprising 49. The maximum step size was taken as 1 unit. The maximum number of evolutions in each memeplex is 6, the step length coefficient is 2, and the maximum number of iterations is restricted to 200. Considering these parameters of SFLA, the optimization model is developed to generate the solutions.

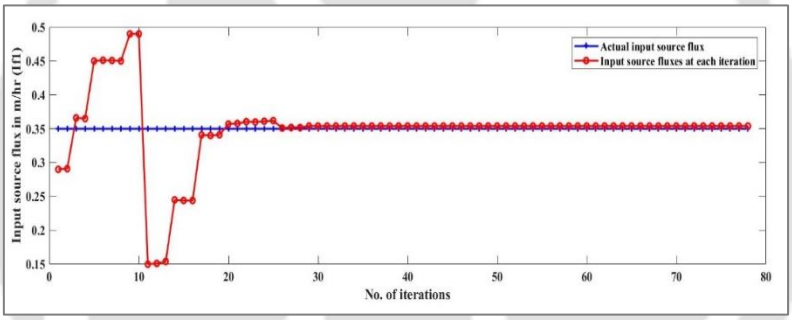
5.4.1.3 Performance of Model 1:

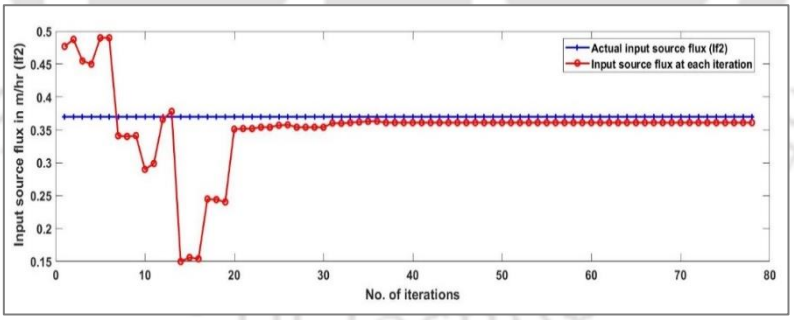
This scenario has six decision variables based on the number of sources ($n=2$). The optimization model will provide the candidate solution to the numerical simulation model to obtain the simulated virus concentration. This candidate solution will be modified based on the objective function value. The process is repeated until the optimal solution is obtained. Table 5. 1 shows the Actual and estimated value along with the relative percentage error using the Numerical Transport Simulation Model (NSM) and SFLA. Figure 5.4 represents the solution plot versus the number of iterations generated by the SFLA algorithm. Initially, the fitness of the solutions is low, and it fluctuates as the algorithm explores the solution space. However, as the number of iterations increases, the algorithm converges on the optimal solution, and the fitness of the solutions increases and eventually stabilizes. Overall, the plot of the solution versus the number of iterations obtained using the SFLA algorithm provides a visual

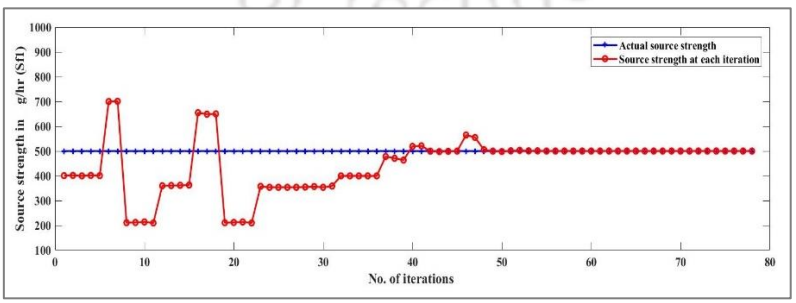
representation of the optimization process and can be used to evaluate the performance of the algorithm with other optimization algorithms.











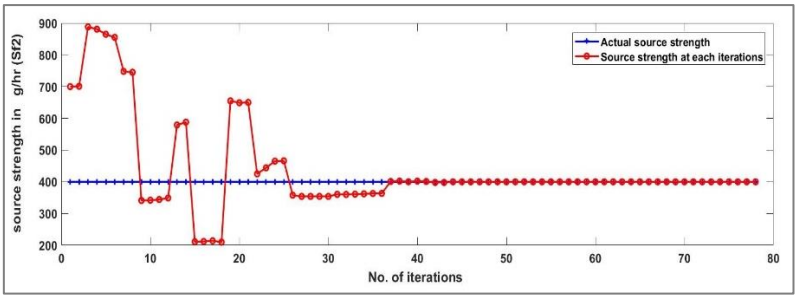


Figure 5. 4: Input source flux (*If*), Source locations (S*l*), and source strength (*Sf*) with No. of iteration for Model 1.

Table 5. 1: Comparison between the estimated and actual virus sources in Model 1

	Lower bound	Upper bound	Actual	Predicted	Relative error (%)	Function value
Sl_1	1	400	84	84	-	2.04E-05
Sl_2	1	400	128	128	-	
If_1	0.1	0.5	0.35	0.354	1.142	
If_2	0.1	0.5	0.37	0.361	2.43	
Sf_1	100	1000	500	500.2	0.04	
Sf_2	100	1000	400	399.92	0.02	

From the analysis, it is found that the model could accurately predict the source location that is Sl_1(4,5,1) and Sl_2 (8,7,1) with 500.2 g/h and 399.92 g/h as virus source strength and 0.354 m/h and 0.361 m/h as the input source flux respectively. The objective function value is calculated, and it is equal to 2.04E-05. This value of the objective function emerges as the best function value; further, no improvement is seen in it. Thus, the function tolerance (stopping criteria) compels the simulation to end at 86 iterations. Based on the relative error calculated using equation (5.7), the error ranges from 0.02 % to 2.43%, which is acceptable. Thus, to check the performance efficiency of the model, this SFLA-based source identification model for scenario 1 is used to identify the sources with different sets of observed data by implementing other measurement errors. The observation obtained from this analysis is discussed in section 5.4.1.4.

5.4.1.4 Effect of different noise levels on source location and fluxes for Model 1:

Concentration measurements cannot be performed accurately in the field every time. Model performance has been evaluated even when the observed data is inaccurate. As a result, the observed concentration data are subjected to measurement errors and analyzed accordingly. Errors are calculated using the normal distribution with sigma (σ) and zero means (μ= 0). Noise

levels (σ) introduced in this study are 0.05, 0.1, 0.15, and 0.2. Table 5. 2 presents the solution obtained using SFLA-NSM after considering different noises in the observed data.

Table 5. 2: Identified source location, source flux, and inflow flux at different noise levels

Number of pollution sources		Actual pollution source location	Estimated source location and fluxes at different noise level				
			$\sigma = 0.0$	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.15$	$\sigma = 0.2$
Sl_1	*Grid ID*	84	84	84	84	84	84
	(x,y)	(4,5)	(4,5)	(4,5)	(4,5)	(4,5)	(4,5)
Sl_2	*Grid ID*	128	128	128	128	128	128
	(x,y)	(8,7)	(8,7)	(8,7)	(8,7)	(8,7)	(8,7)
Sf_1		500	500.2	498.66	499.29	524.88	409.23
Sf_2		400	399.92	402.25	400.99	366.35	444.25
If_1		0.35	0.354	0.356	0.352	0.372	0.319
If_2		0.37	0.351	0.37	0.3669	0.356	0.352

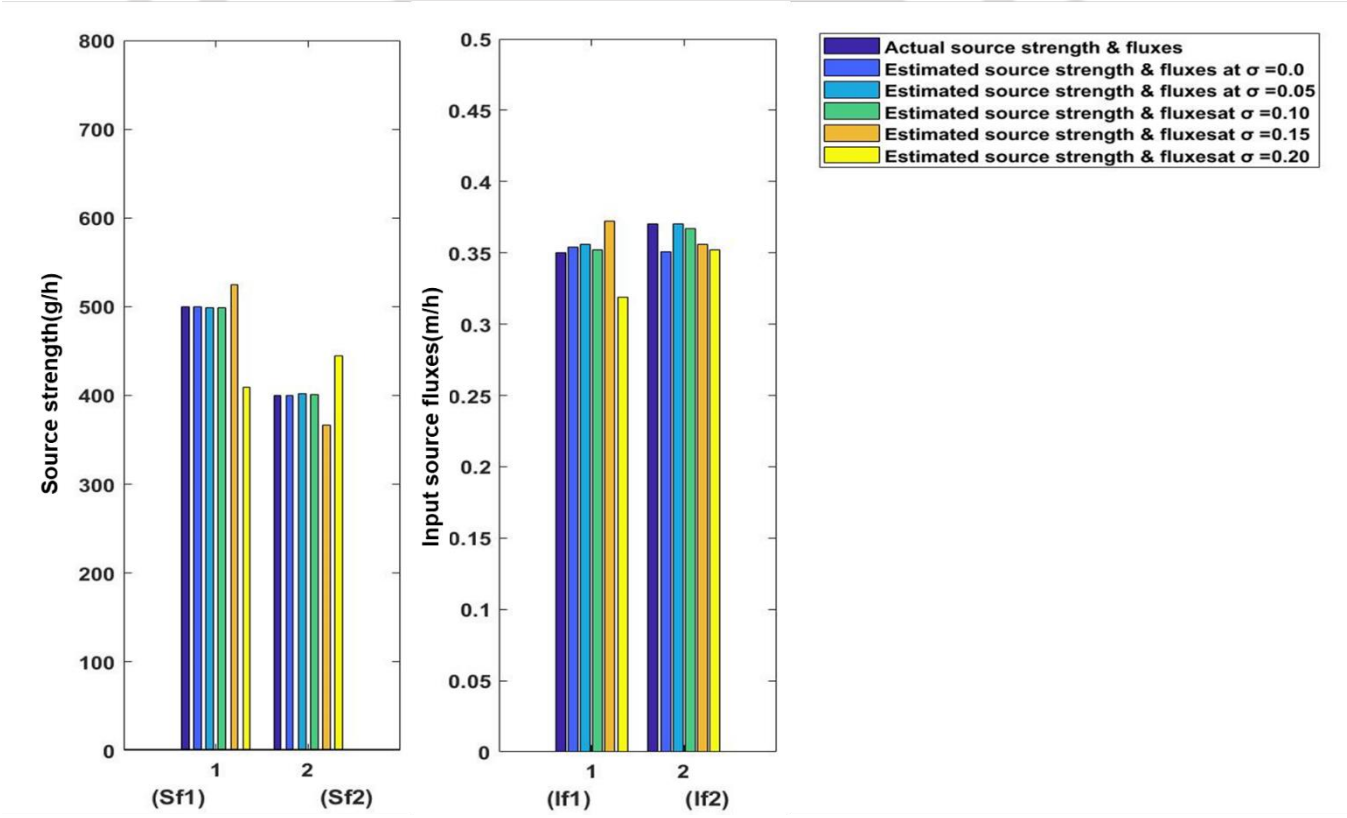


Figure 5.5: Comparison of the Relative error (%) estimated and actual fluxes obtained at different noise level.

Table 5.3: Comparison of the Relative error (%) estimated and actual fluxes obtained at different noise levels.

Number of pollution sources	Actual pollution source location	Relative error (%)				
		$\sigma = 0.0$	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.15$	$\sigma = 0.2$
Sf_1	500	0.04	0.268	0.142	4.97	18.20
Sf_2	400	0.02	0.5625	0.2475	8.41	11.062
If_1	0.35	1.1428	0.5714	1.142	6.2	8.85
If_2	0.37	5.1351	5.1351	0.8378	3.78	4.86

From Table 5.3, it is observed that by adding various noise levels to the iterative search, the present model can identify the exact locations of the sources. Figure 5.5 depicts the inflow source fluxes and the source strength estimated for two locations with different noise levels. Based on the observation, the estimated source flux and strength at the two locations are close to the actual value at different noise levels. Although no trend has been observed with the inflow of source fluxes with increasing noise levels, an increase in source strength is observed with the rise in noise level. Referring to Table 5.3, it is seen that the predicting performance of the model is satisfactory up to σ = 0.1, but when the noise level is increased, the model could predict up to four variables, but the remaining two variables provide non-unique solutions. The relative error obtained from the estimated source fluxes and strength at different noise levels is calculated and listed in Table 5.4. From the investigation, it can be concluded that the proposed methodology could accurately predict the source location, inflow source fluxes, and source strength up to σ = 0.1; after that, the model provides nonunique optimal solutions. Thus, the source identification model formulated for two-dimensional search space can be applied to identify *n* number of sources presented in the XY plane when the depth of the source is known.

5.4.2 Virus source identification model in 3D-space (Model 2):

In the first scenario, we consider virus sources in the top layer of the unsaturated zone. But there may be some situations where the virus sources may present both in the saturated and unsaturated zone. In such a case, we cannot use the source identification model developed in the first scenario; instead, we must search the source location in three-dimensional space.

5.4.2.1 Conversion of 3D domain to 1-D domain:

To simplify the complexity of variables in a three-dimensional search space, the XYZ coordinates are transformed into a single dimension by associating grid IDs with each grid. This method effectively combines the three coordinates into a single reference point. This

approach allows for the identification of grid IDs instead of dealing with the task of identifying the coordinates for a given '***n***' number of sources. The 3D- search spaces are marked with grid IDs using the equation (5.9).

$$IDs = [(y-1)Ny + x] + [(z-1) \times Nx \times Ny] \quad (5.9)$$

Here, *Nx*, *Ny*, and *Nz* are the number of grids along the x-axis, y-axis, and Z-axis. *x*, y, *and* z is the coordinate along the 3D domain.

5.4.2.2 When one source location is present in an unsaturated zone and another in a saturated zone (Model 2):

To develop the virus source identification model in three-dimensional space, the example problem developed for Scenario 2 in Chapter 3 is considered. The study area showing the source location is presented in Figure 5.6. The model's grid distribution along with the source location, pumping well, and observation well is shown in Figure 5. 7. In this study area, one source location is present in the unsaturated zone at coordinates *Sf1* (8,16,1) and another in the saturated zone *Sf2*(10,13,8). Using equation (5.9), the grid IDs are assigned such that *Sf1*= 308 and *Sf2*= 3050. So now we need to identify the source location at *Sf1*= 308 and *Sf2*= 3050. Since the number of grids along x,y and z direction are *Nx*= 20, *Ny*=20, and Nz =40 so the total number of grids is 16000. As such, the model must detect the virus source location between 1 and 16000.

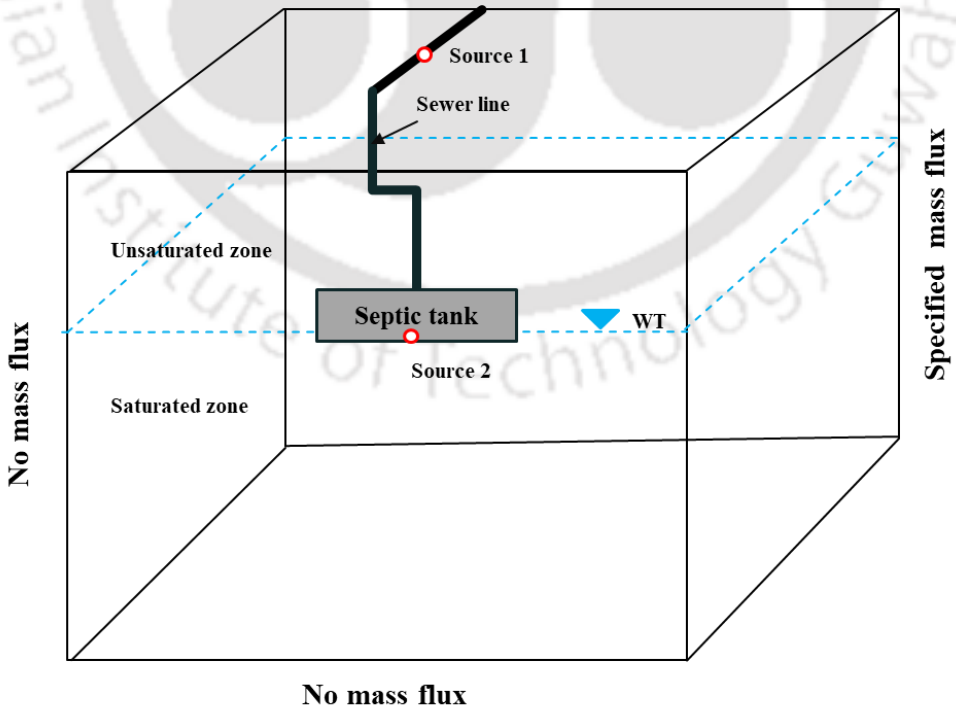
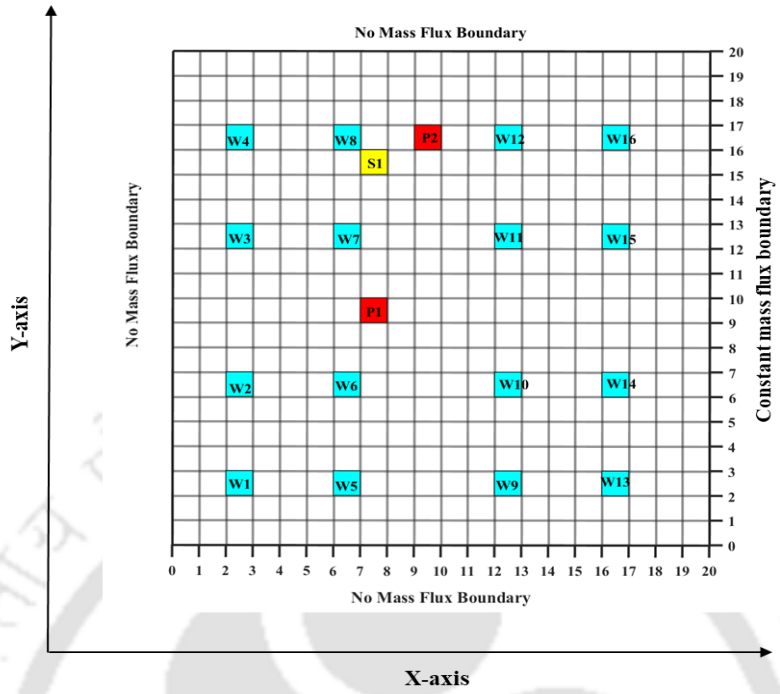
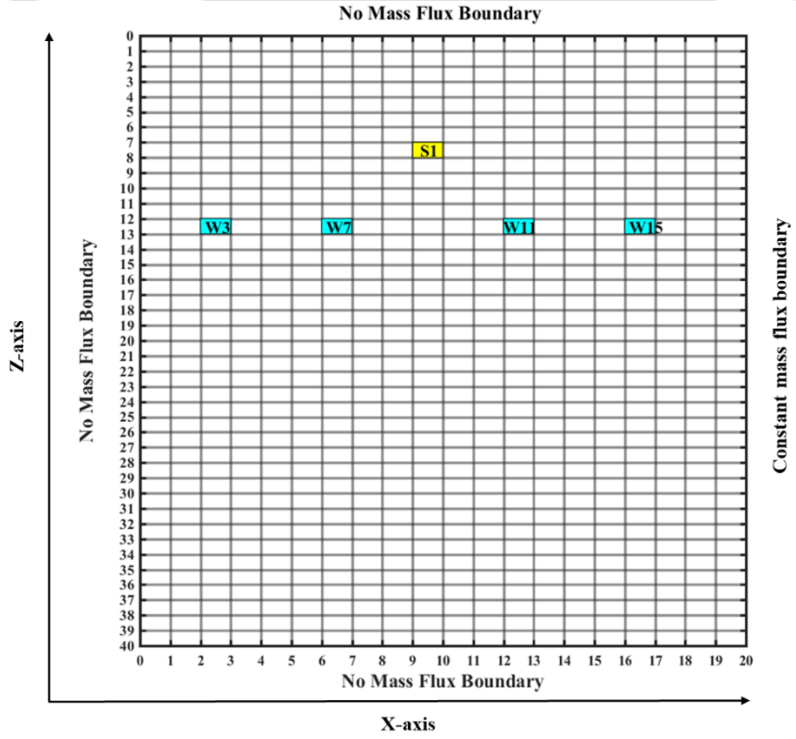


Figure 5.6: Graphical representation of the synthetic model that is considered for Scenario 2



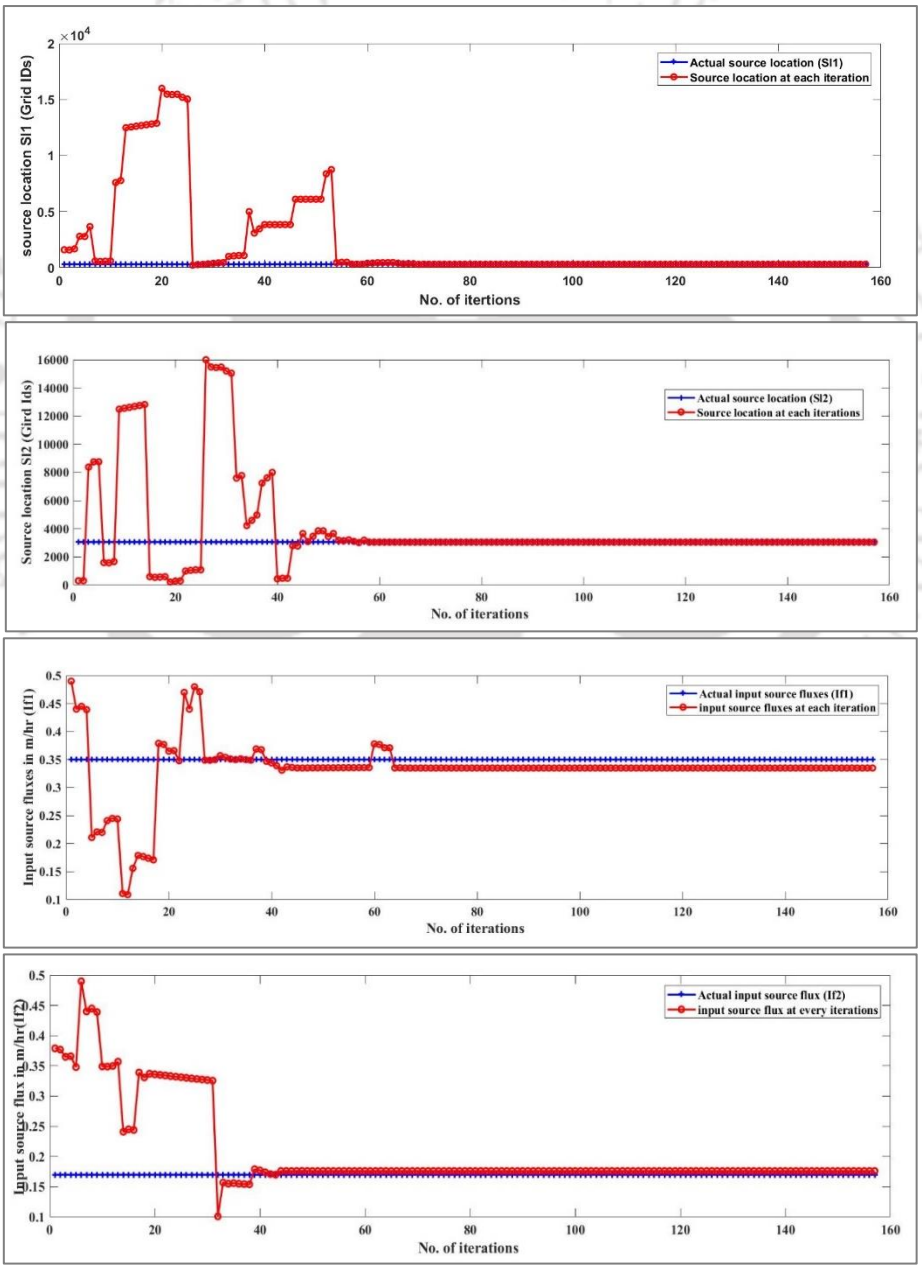
(a) Grid representation of the top view of the Model (Scenario2) along X-Y plane



(b) Grid representation of the sectional view of the Model (Scenario 2) along X-Z plane

Figure 5. 7: Grid distribution of the model showing the source location, pumping well and observation

The source identification model is developed after converting the three-dimensional space to one dimension, providing grid IDs using equation 5.9. In this scenario, the parameter used for SFLA is similar to that used in scenario 1. The model solves the objective function in equation 5.1 to obtain the source location, input fluxes, and source strength. The model uses the numerical simulation model developed in Chapter 3 of Scenario 2. The model is coupled with SFLA to obtain optimal solutions. The model has six decision variables, which are two input source flux (*If*), two source locations (*Sl*), and the corresponding source strength (*Sf*). Table 5.4 shows the actual and estimated value and the relative error obtained from the source identification model developed for scenario 2.



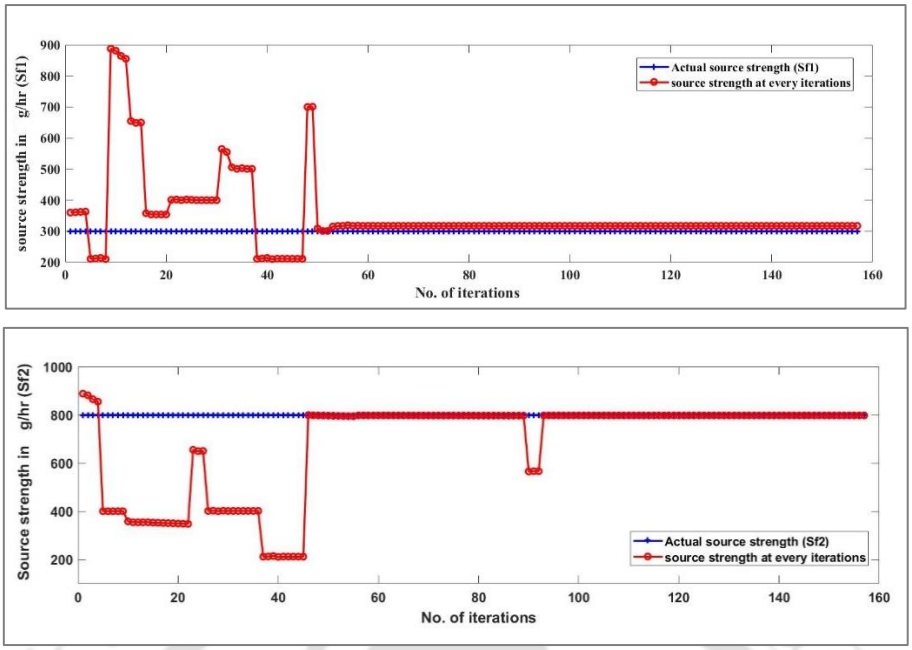


Figure 5. 8: Input source flux (If), Source locations (Sl), and source strength (Sf) with No. of iteration for Scenario-2

Figure 5. 8 shows the variation of input source flux (*If*), source locations (*Sl*), and source strength (*Sf*) with the number of iterations obtained using the source identification model. The figure shows how the quality of the solutions generated by the SFLA algorithm changes over time as the number of iterations increases. As the number of iterations increases, the SFLA algorithm generates new sets of input parameters and evaluates their fitness. The best solutions from each iteration are retained and used to generate new solutions in the next iteration. This iterative process continues until the desired level of accuracy or convergence is achieved.

Table 5.4: Result II: Estimated virus source location with relative error obtained using SFLA in Scenario 2.

	Lower bound	**Upper bound**	**Actual**	**Predicted**	**Relative error (%)**	**Function value**
Sl_1	1	16000	308 (8,16,1)	308 (8,16,1)	--	6.89E-04
Sl_2	1	16000	3050 (10,13,8)	3050 (10,13,8)	--	
If_1	0.1	0.5	0.35	0.335	4.2857	
If_2	0.1	0.5	0.17	0.176	3.5294	
Sf_1	100	1000	300	317.21	5.7367	
Sf_2	100	1000	800	797.66	0.2925	

The solutions obtained from the linked simulation-optimization model to identify the virus contaminant source in three-dimensional space is presented in Table 5.4. The analysis shows that the estimated source locations are (8,16,1) and (10,13,8) with 317.21 g/h and 797.66 g/h as virus source strength and 0.335 m/h and 0.176 m/h as the input source fluxes, respectively. The objective function value is calculated and is equal to 6.89E-04. This objective function value emerges as the best function value, as no improvement is seen in the analysis. Thus, the function tolerance (stopping criteria) compels the simulation to end at 157 iteration steps. Based on the relative error calculated using equation (5.7), the error ranges from 0.2925% to 5.73%, which is acceptable. Thus, to check the performance efficiency of the SFLA-based source identification model in three-dimensional space, the virus contaminant sources were identified with different sets of observed data in conjunction with measurement errors. The observation obtained from this analysis is discussed in section 5.4.2.3.

5.4.2.3 Effect of different noise levels on source location and fluxes for Model 2:

In this study, the model performance has been evaluated even when the observed data is inaccurate. As a result, the observed concentration data are subjected to measurement errors and analyzed accordingly. Errors are calculated using the normal distribution with sigma (σ) and zero mean (μ= 0). Noise levels (σ) introduced in this study are 0.05, 0.1, 0.15, and 0.2. Table 5. 5 provides the estimated source locations, input source fluxes, and source strength at different noise levels.

Table 5. 5: Identified source location, source flux, and inflow flux at different noise levels for Model 2

Number of pollution sources	**Actual pollution source location (*IDs*)**	**Estimated source location (*i,j*)**				
		At zero noise level	**At noise level σ = 0.05**	**At noise level σ = 0.1**	**At noise level σ = 0.15**	**At noise level σ = 0.2**
Sl_1	308	308	308	308	308	308
Sl_2	3050	3050	3050	3050	3050	3050
If_1	0.35	0.335	0.351	0.366	0.310	0.374
If_2	0.17	0.176	0.171	0.150	0.140	0.237
Sf_1	300	317.21	296.98	338.22	499.56	338.23
Sf_2	800	797.66	811.29	800.12	586.004	841.77

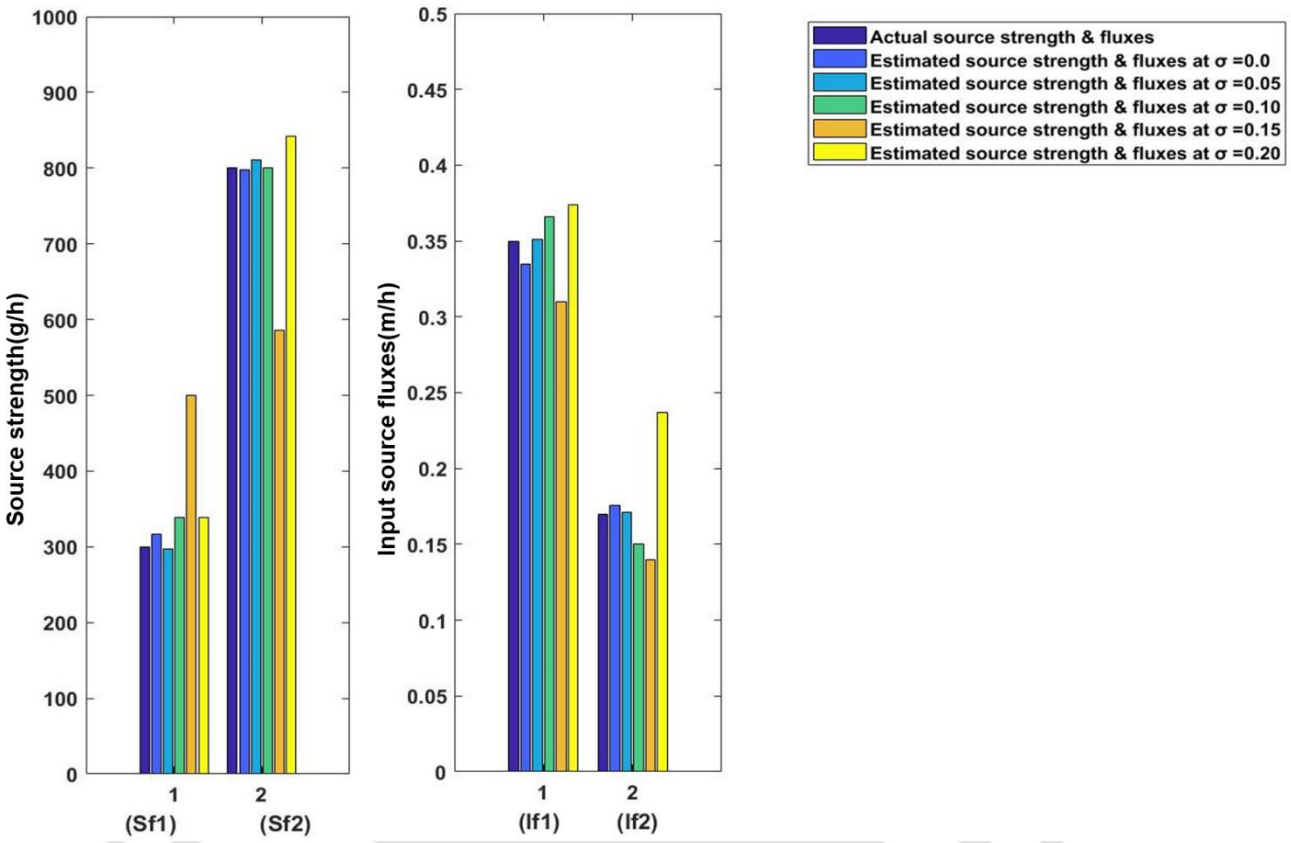


Figure 5. 9:Comparison of the relative error (%) estimated fluxes at different noise levels for scenario 2.

Table 5. 6: Comparison of the Relative error (%) estimated and actual fluxes obtained at different noise levels for scenario 2:

Number of pollution sources	Actual pollution source location	Relative error (%)				
		σ = 0.0	σ = 0.05	σ = 0.1	σ = 0.15	σ = 0.2
Sf_1	0.35	4.2857	0.2857	4.5714	11.4285	6.8571
Sf_2	0.17	3.5294	0.5882	11.76471	17.6470	39.4117
If_1	300	5.7367	1.0066	12.74	66.52	12.7433
If_2	800	0.2925	1.4112	0.015	26.7495	5.2212

From Table 5. 5, it is observed that by adding various noise levels to the iterative search, the present model can identify the exact locations of the source. However, the input source fluxes (*If*) and virus source strength (*Sf*) yield a different result. When the noise level is zero, the estimated source fluxes and source strength closely resemble the actual ones, but when the noise level is increased, there appears to be some divergence between the estimated and the actual fluxes. Referring to Table 5. 5, the estimated input source flux ranges from 0.310 m/h to

0.374 m/h for source If_1, and for source If_2, it ranges from 0.140 m/h to 0.237m/h. The virus source strength (Sf_1) ranges from 296.98 g/h to 499.56 g/h; also, for Sf_2, it ranges from 586.004 g/h to 797.66 g/h. Although no trends have been observed with the inflow and source fluxes with increasing noise levels, the relative error is calculated to understand more about the predicted performance. Table 5. 6 provides the relative error between the actual and estimated virus sources at different noise levels in observed data. The analysis up to 0.1 noise level shows that the model could estimate the near-optimal solution with a minimum error within the acceptable range.

5.5 Conclusion:

The study developed a model for identifying virus sources in coupled unsaturated-saturated groundwater aquifers using a linked simulation-optimization approach. This minimizes the error between observed and simulated virus concentrations, integrating a three-dimensional virus transport model with the Shuffled-Frog Leaping Algorithm. The model concludes as follows:

- Two virus source identification model is developed: (a) To identify the source location in an unsaturated zone in a two-dimensional search space and (b) To identify the source location in both saturated and unsaturated zones in a three-dimensional space. The source identification model was developed by simplifying the model into one-dimensional space to reduce computational complexity.
- The findings indicated that the SFLA-based virus source identification model can effectively identify both models' sources and pathways of virus contamination.
- This study developed a source identification model using observed data with varying noise levels. The model accurately predicted virus contamination locations at different noise levels. However, when estimating virus source flux and strength, the model could tolerate up to 0.1 noise in the observed data. The identified noise level in the data significantly influenced result accuracy and reliability.
- In conclusion, the developed source identification model can identify ***n*** number of sources in both unsaturated and saturated zones. This model can be developed after the number of source locations is estimated using the model available in the literature.

CHAPTER 6

IMPROVED SOURCE IDENTIFICATION MODEL INCORPORATING RNN, NUMERICAL TRANSPORT SIMULATION MODEL (NSM) AND SFLA

6.1 Overview

This chapter presents an advanced approach to identify virus sources in groundwater aquifers. The chapter introduces a novel model that combines three key components: the Recurrent Neural Network (RNN), the Numerical Transport Simulation Model (NSM), and the Shuffled Frog-Leaping Algorithm (SFLA). This chapter is mainly about improving the source identification model developed in Chapter 5. Initially, the model is developed to improve the computational time efficiency using Recurrent Neural Network(RNN) and SFLA. Later, this chapter proposed a new methodology using the RNN- NSM –SFLA for achieving efficiency in computational time and predicting performance in identifying the virus sources. Further, the chapter summarizes the findings and contributions of the proposed improved source identification model. Additionally, it suggests avenues for future research and further enhancements to improve the accuracy and applicability of the model.

6.2 Introduction:

Modeling or managing contaminant transport at the field scale requires consideration of contaminant sources. In any groundwater management problem, it is essential to identify contaminant sources and fluxes in aquifers. An inverse optimization model was used to develop a source identification model in the previous chapter, and the results demonstrated that this model could accurately identify the source locations and fluxes. Based on computational time efficiency, the virus source identification model developed so far is computationally expensive. This is due to the incorporation of the coupled flow and virus transport model, with the optimization model, which makes the model highly nonlinear. Consequently, the model becomes highly expensive as a whole. An approximate simulation model and optimization model needs to be used to overcome this problem. Even though computational time is reduced, the predictive performance of such model will decrease. In other words, the solution achieved may differ from the global optimal solution. However, it may be close to the optimal solution. Although incorporating a sophisticated numerical simulation model will improve its performance, but its computational time will increase. To achieve efficiency in computational

time and predicting performance, a new methodology using the ANN, MODFLOW and MT3DMS and the GA model was developed to identify pollution sources (Borah & Bhattacharjya, 2016). They initially generated the function value with the population set generated by GA using ANN. They sorted the population into two groups, better x% and rest (1− x)%. Later, the fitness value of the better x% of the population is calculated using the GMS model to identify the sources. This model has few assumptions and limitations-(a) Since the model evaluates the fitness value with the x% of the population again using GMS, the number of function evaluations becomes more (b) Also, the model assumed that the source location of the contaminant is known and it is applicable to identify the source strength.

Considering the above-mentioned limitations, a new methodology has been introduced using the Numerical Transport Simulation Model - RNN models in conjunction with the Shuffled Frog-Leaping Algorithm (SFLA) to estimate the pollution sources. This modified simulation-optimization model initially involves calculating the fitness values using the 2000 input-output data generated using a Numerical Transport Simulation Model (NSM). The solutions are arranged in two groups based on the fitness value: the better solutions in x% and the remaining (100-x%) in the other group. When the candidate solutions generated by the optimization model at each iteration fall within a x% group, indicating a high fitness and likelihood of being a pollution source, the NSM model is employed. Otherwise, the RNN is utilized for further simulation and optimization. This modified approach offers several advantages. (a) Firstly, it significantly reduces the number of function evaluations as the function value is calculated initially rather than calculating at each set of population size. (b) This model tried to estimate the source locations along with the source fluxes. This methodology thus helps to achieve faster results and makes the source identification process more practical for real-world scenarios. Secondly, by integrating both numerical simulation and RNN models, the modified model benefits from the strengths of each. Furthermore, the modified model is designed to obtain a globally optimal solution by combining the SFLA optimization algorithm with the fitness values calculated using the NSM. This enhances the model's ability to identify virus sources accurately.

6.3 RNN-SFLA-based virus source identification model (Model 1).

Chapter 5 shows that the source identification model can be developed considering source location in both saturated and unsaturated zones for two-dimensional and three-dimensional search space. However, the analysis took around 1 day, 5 hours, 45 minutes, and 10 seconds to complete the whole analysis for scenario 1, and for scenario 2, it took around 1 day, 16 hours,

24 minutes, and 39 seconds, which is very time-consuming. This model was run in computer system with specification of processor: Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz 3.41 GHz, 8GB RAM. Thus, a source identification model is developed initially using an approximate simulator, Recurrent Neural Network (RNN) to reduce the computational time. Instead of using the Numerical Transport Simulation Model (NSM) previously developed for coupled flow and virus transport in groundwater aquifers, the RNN models are employed as a replacement. This RNN model is then linked with the optimization model to identify unknown virus sources. This combination of the RNN model and SFLA offers a more efficient and time-effective approach for source identification in the context of virus transport in groundwater systems.

6.3.1 Methodology:

In the RNN model, the inputs consist of the source locations (*Sl*), source strength (*Sf*), and inflow source flux (*If*), while the outputs are the virus concentration values at different observation wells and time steps. Specifically, the RNN model takes in these input parameters and generates predictions for the virus concentration at various observation wells over time. The number of RNN models developed equals the number of observation wells. Once the RNN models are developed and trained using the input-output patterns, it is called upon by the optimization model to calculate the simulated virus concentration. The decision variables are the input parameters, and the total number of variables is based on the number of sources (*n*). This model uses Shuffled Frog-Leaping Algorithm (SFLA) as an optimization algorithm. The objective function will minimize the error between the simulated concentration obtained from the RNN model and the observed concentration which is given by equation (6.1). Figure 6. 1 shows the methodology proposed to develop the source identification model using RNN-SFLA.

$$\text{Minimize } f_x = \sum_{j=1}^{M}\sum_{i=1}^{N}\left|OC_i^j - SC_i^j\right| \tag{6.1}$$

$$\text{Subject to } C_W = f\left(If, Sl, Sf\right) \tag{6.2}$$

$$If_{\min} \le If \le If_{\max} \tag{6.3}$$

$$Sl_{\min} \le Sl \le Sl_{\max} \tag{6.4}$$

$$Sf_{\min} \le Sf \le Sf_{\max} \tag{6.5}$$

Here, f_x is the objective function for the present optimization model with x number of virus sources; OC_i^j is the observed concentration at i^{th} time step for j^{th} well location; SC_i^j is the simulated concentration at i^{th} time step for j^{th} well location obtained from RNN model or numerical simulation model; $M$is the total number of observation wells and; $N$is the total number of time steps; $C_W$is the concentration vectors of the simulated concentration; If represents the inflow flux applied in the groundwater flow model such that $If = [If_1, If_2 \ldots If_x]^T$; Sf represents the virus source fluxes such that $Sf = [Sf_1, Sf_2 \ldots Sf_x]^T$; Sl represents the virus source location vectors such that $Sl=[Sl_1, Sl_2 \ldots\ldots Sl_x]^T$; The suffix '$min$' and '$max$ ' represent the source flux's lower and upper bounds, source location, and inflow flux in the source identification.

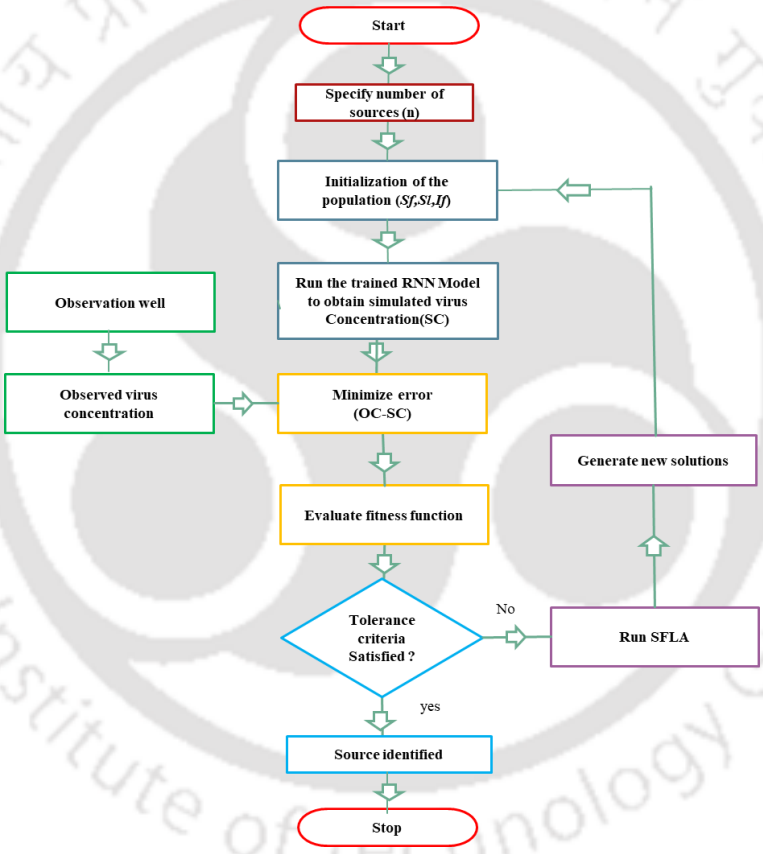


Figure 6. 1: Flowchart showing the RNN-SFLA-based source identification model

6.3.2 Recurrent Neural Network (RNN):

Recurrent Neural Networks (RNNs) are a type of neural network designed to handle sequential data by incorporating a hidden layer with a hidden state. The hidden state in RNNs serves as a memory that allows the network to retain information about previous steps or events in the sequence. In traditional neural networks, inputs and outputs are treated as independent entities, but for tasks that involve sequential data, such as natural language processing or time series

analysis, the context and dependencies between the elements in the sequence are essential. RNNs address this by introducing connections that allow information from previous steps to influence the computation in the current step. One notable aspect of RNNs is that they share the same set of parameters across all steps or hidden layers. This parameter sharing reduces the complexity of the model compared to other neural network architectures, where each input has its own set of parameters. This shared parameter structure allows RNNs to process sequences of varying lengths effectively and generalize well to unseen data.

6.3.2.1 Architecture of Recurrent Neural Network (RNN):

The architecture of the RNN model described in Figure 6. 2 consists of three main layers: the input layer (I), the hidden layer (Z), and the output layer (O). Each layer plays a specific role in processing the input and producing the desired output: the virus concentration in this case. The input layer (I) is where the source location (*Sl*), source fluxes (*Sf*), and input velocity fluxes (*If*) are fed into the model. These inputs provide the necessary information for the RNN model to predict the virus concentration. The values of these inputs are used as the initial input for the RNN model. The hidden layer (Z) considers the inputs from the input layer (I) and the hidden state from the previous time step. This enables the RNN model to incorporate both the current input and the information from the past to make predictions. The hidden layer performs computations using the RNN model's parameters to update its hidden state and produce an output. The output layer (O) takes the hidden state of the current time step and computes the predicted virus concentration. The output layer's activation function produces the final output, representing the virus concentration at a specific time or location.

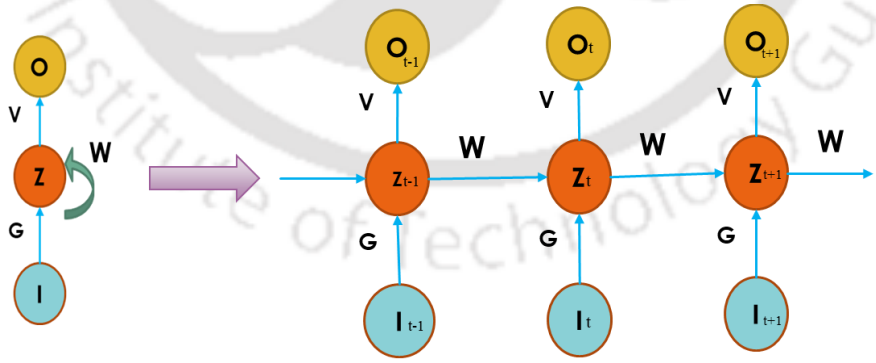


Figure 6. 2: Architecture of (a) folded RNN model and (b) unfolded RNN.

The hidden state is updated using the following relations:

$$Z_t = f\left(Z_{t-1} + I_t\right) \tag{6.6a}$$

$$Z_t = \tanh\left(V_{zz} Z_{t-1} + V_{Iz} Z_t\right) \tag{6.6b}$$

$$O_t = \left(V_{zo} Z_t\right) \tag{6.7}$$

Where, Z_t is the current hidden state, O_t is the output, and I_t is the input vector at a time. The input, hidden, and output weights are represented by G, W, and V, respectively. V_{zz} is the weight at recurrent neuron, V_{Iz} is the weight at input neuron and the V_{zo} is the weight at the output.. The hyperbolic tangent activation function is noted as *tanh*.

6.3.2.2 Transfer function and optimization algorithm adopted for the RNN model:

The selection of an optimization algorithm is indeed an important step in training an RNN model. In the case of the described study, the '*trainscg*' training function available in MATLAB has been utilized for training the current RNN model. The '*trainscg*' function employs the scaled conjugate gradient method for updating the weights and biases of the RNN model. This method is based on the principle of conjugate gradients, an optimization technique used to find the minimum of a function. To achieve the best performance in terms of mean square error (MSE), computation time, and iteration time, the combination of the *'tansig-purelin*' transfer functions for the hidden and output layers was effective. The '*tansig*' transfer function generates an output ranging from -1 to +1, while the '*purelin*' transfer function produces a linear output. This combination allows for the appropriate activation of the hidden and output layers in the RNN model. A second hidden layer with a different transfer function combination was also introduced to further enhance the model's performance. In this case, the *'tansig-logsig'* transfer functions were employed. The '*logsig*' transfer function, also known as the logistic sigmoid function, maps inputs to outputs ranging between 0 and 1. The combined transfer functions and the '*trainscg*' training function were chosen due to their computational efficiency and good convergence properties. The '*trainscg*' function and the selected transfer function provide a practical optimization approach for training the RNN model and achieving the desired performance metrics.

6.3.2.3 Evaluation of RNN model performance using statistical criteria:

To determine the performance of the Recurrent Neural Network (RNN) models, three different statistical parameters have been determined, i.e., Root Mean Square Error (RMSE), Average Absolute Relative Error (AARE), and Coefficient of Determination (R^2). The equation to calculate this error is given by the equation (6.8) to (6.10)

MSE can be estimated using the following equation,

$$AARE = \frac{1}{n_c} \sum_{n_c=1}^{n_c} \left| \frac{OC_i^j - SC_i^j}{SC_i^j} \right| \times 100 \tag{6.8}$$

$$RMSE = \sqrt{\frac{1}{n_c}\sum_{n_c=1}^{n_c}\left(OC_i^j - SC_i^j\right)^2} \tag{6.9}$$

$$R = \frac{\sum_{n_c=1}^{n_c}\left(OC_i^j - \overline{OC_i^j}\right)\left(SC_i^j - \overline{SC_i^j}\right)}{\sqrt{\sum_{n_c=1}^{n_c}\left(OC_i^j - \overline{OC_i^j}\right)^2\left(SC_i^j - \overline{SC_i^j}\right)^2}} \tag{6.10}$$

Where, OC_i^j is the observed virus concentration at well location i^{th} for j^{th} time steps obtained using numerical virus transport model; SC_i^j is the simulated virus concentration at well location i^{th} for j^{th}time steps obtained using RNN model; $\overline{SC_i^j}$ is the mean of the simulated virus concentration; $\overline{OC_i^j}$ is the mean of the observed virus concentration; n_c is the total number of concentration data. When the value of RMSE, AARE is small, then it considers that the solution obtained from the RNN model is significant. Also, if the value of R^2 is close to 1, then it means that the simulated concentration obtained from the RNN model has a close relationship with the observed virus concentration.

6.3.3 Observation from source identification model using RNN-SFLA (model 1):

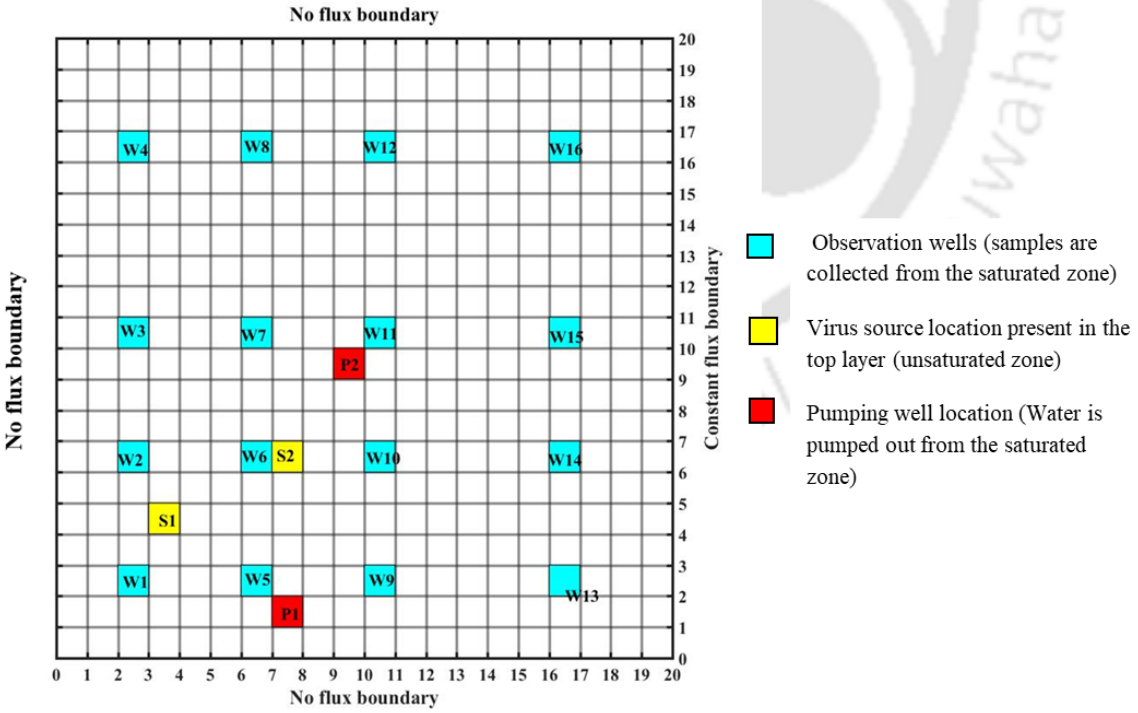


Figure 6. 3: Study area showing the actual source locations, observation well locations and pumping wells.

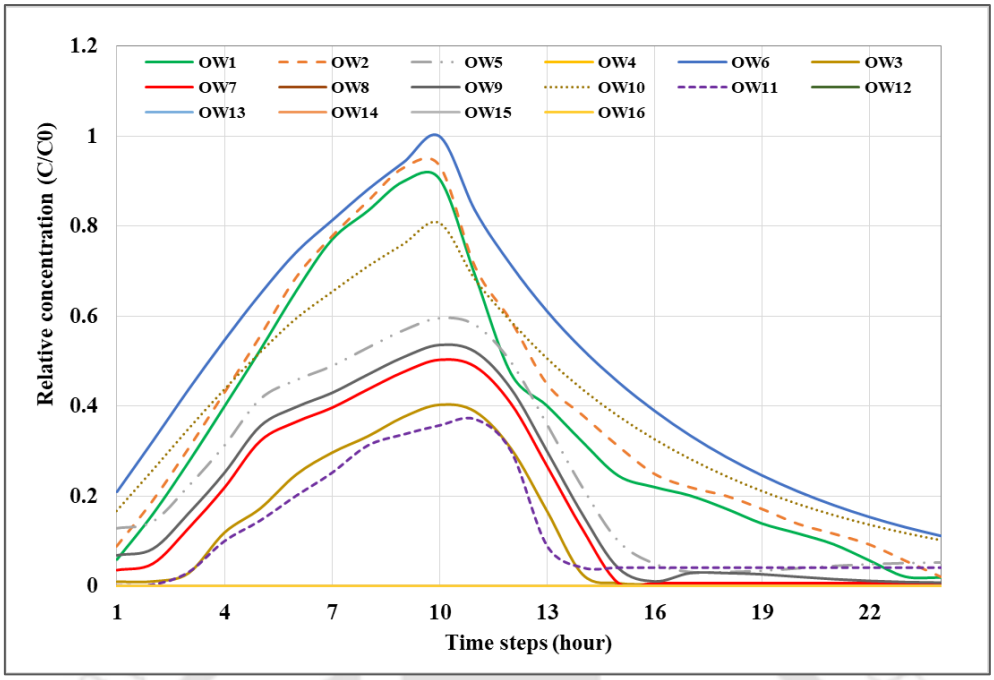


Figure 6. 4: Breakthrough curve of virus concentration for 16 observation wells.

In this study, the Numerical Transport Simulation Model (NSM) is replaced by an approximate simulator developed using the Recurrent Neural Network (RNN) to reduce the computational time. To develop the source identification model using RNN and SFLA, the example problem developed for scenario 1 in Chapter 3 is used.

The location of the virus sources, along with the pumping well and the observation well location, is shown in Figure 6. 3. In this scenario, the no. of source location is known n =2, it also assumes that it is at the top surface of the unsaturated zone. Thus, this model is developed by considering the virus source to be present only at the first layer so the model can search the location in two-dimensional space. For three-dimensional space, we can increase the number of variables, that is, for each source location, there will be three variables for three coordinate axes (x,y,z) and then estimated. Figure 6. 4 shows the variations in the breakthrough curve of virus concentrations over time at 16 observation wells obtained from the NSM developed in chapter 3. To identify the optimal virus source locations, this study will focus on these observation wells for observed concentration and the simulated results obtained from the trained RNN models.

6.3.3.1 Development of the RNN model:

In this study, 16 observation wells are considered, and the virus concentration data obtained from these wells (specifically from the saturated zone) are used as the observed concentration. As mentioned above, RNN models must be equal to the number of observation wells, so 16

RNN models are developed. The RNN model developed in this study has two hidden layers and is designed with a schematic arrangement of 8-40-25-48 neurons. The input layer consists of 8 variables, which include 4 variables for the source location (*Sl*), 2 variables for the source fluxes (*Sf*), and 2 variables for the input velocity fluxes (*If*). The output layer consists of 48 neurons, that is, the virus concentration at each observation location for 24 hours with a time interval of 30 minutes. The number of neurons will vary based on the number of sources. To train and test the RNN models, a dataset of 2000 input-output patterns is generated using the numerical simulation model developed for scenario 1 in Chapter 3. The model took one day 2 hours 10 minutes and 3 seconds to generate the dataset for training the RNN model. Each pattern corresponds to a specific combination of source locations, source strength, inflow source flux as input, and virus concentration values at different observation wells and time steps as the output. Once trained, the model can predict the virus concentration for new input data and thus identify the optimal source locations when linked with the optimization model. A schematic arrangement of the neuron pattern considered while developing the RNN models is shown in Figure 6. 5.

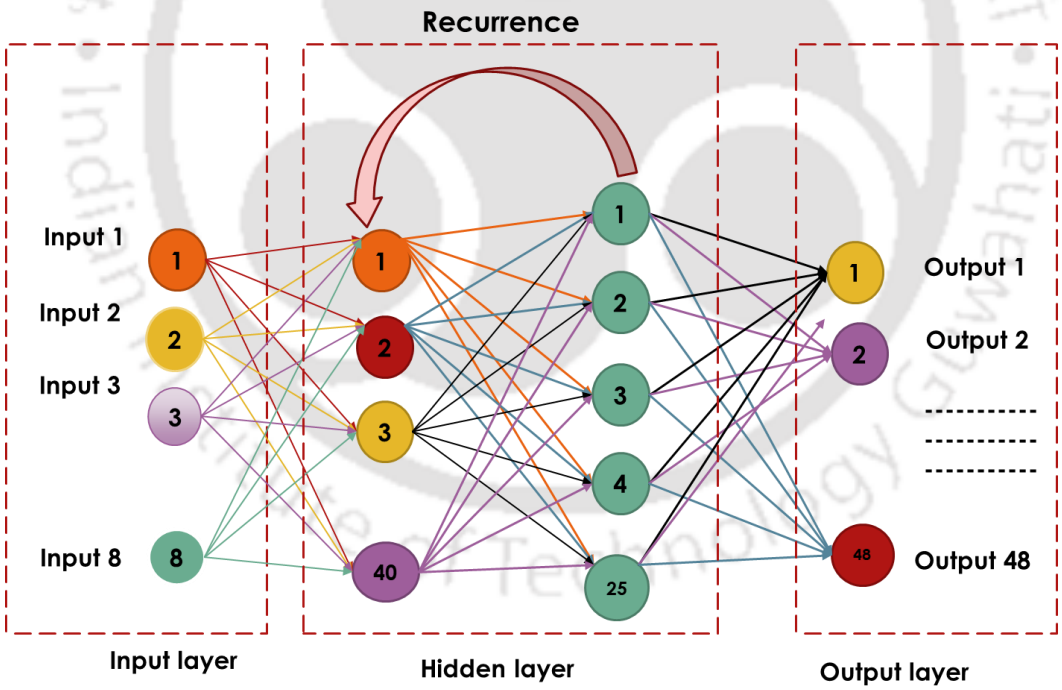


Figure 6. 5:RNN model network.

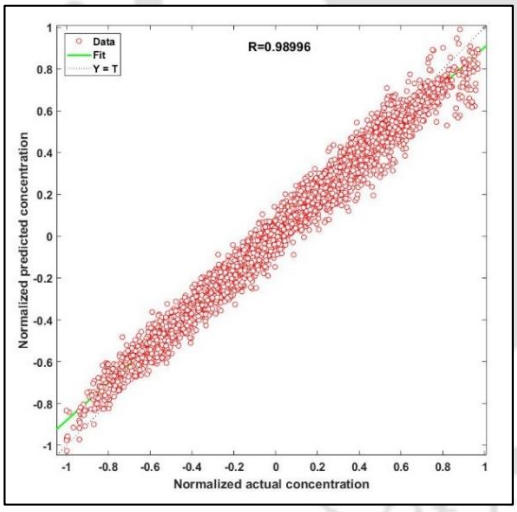
6.3.3.2 Performance of RNN models:

After generating 2000 input-output data using the Numerical Transport Simulation Model (NSM) by coupling the groundwater flow and transport model developed in Chapter 3 for scenario 1, the RNN simulator is trained, tested, and validated. The performance of the trained

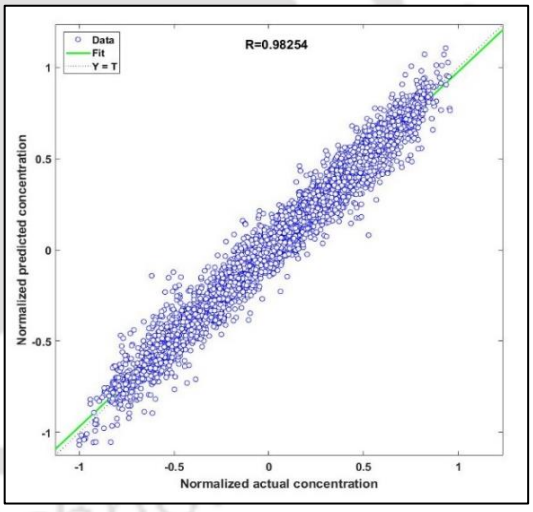
RNN model using AARE (Average Absolute Relative Error) and RMSE (Root Mean Squared Error) is listed in Table 6. 1. These metrics are commonly used to evaluate the accuracy and precision of prediction models. A lower value for both AARE and RMSE indicates better performance of the RNN model. For each observation well, RNN models are developed, and the scatter plots at 48 time steps are plotted as shown in Figure 6.6.

Table 6. 1: Performance of the trained RNN model using AARE and RMSE

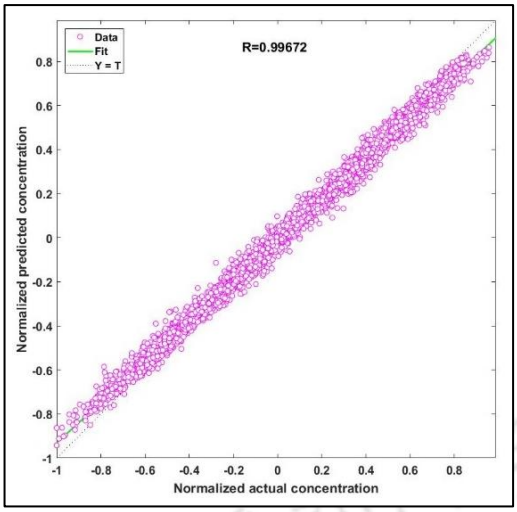
RNN MODEL	AARE	RMSE	ANN MODEL	AARE	RMSE
Well 1	0.77	0.829	**Well 9**	0.91	0.40
Well 2	0.82	0.72	**Well 10**	0.76	0.52
Well 3	0.88	0.18	**Well 11**	0.83	0.16
Well 4	1.98	0.34	**Well 12**	1.86	0.21
Well 5	0.72	0.32	**Well 13**	0.97	0.71
Well 6	0.66	1.22	**Well 14**	0.54	0.62
Well 7	0.98	0.27	**Well 15**	0.28	0.81
Well 8	0.77	0.69	**Well 16**	0.94	0.87



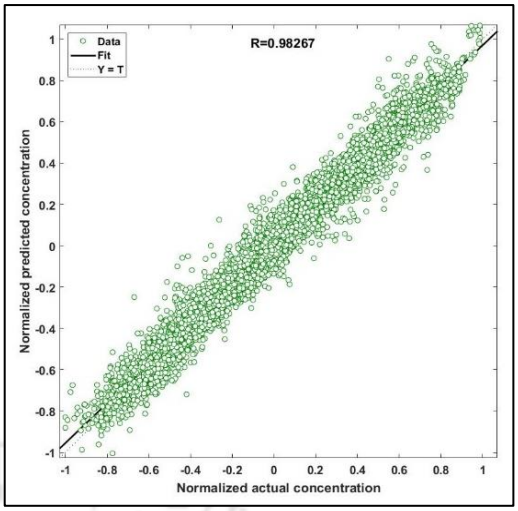
(a) Well 1



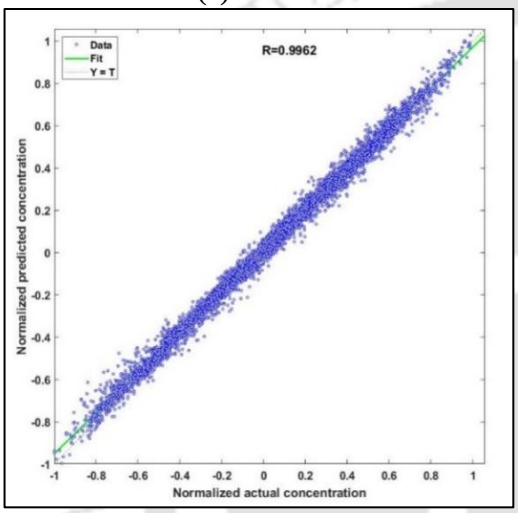
(b) Well 2



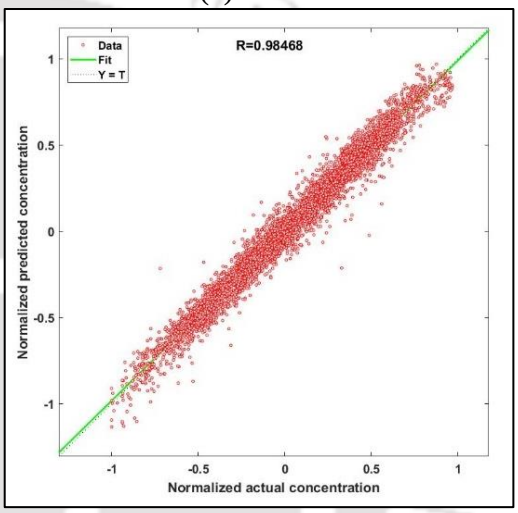
(c) Well 3



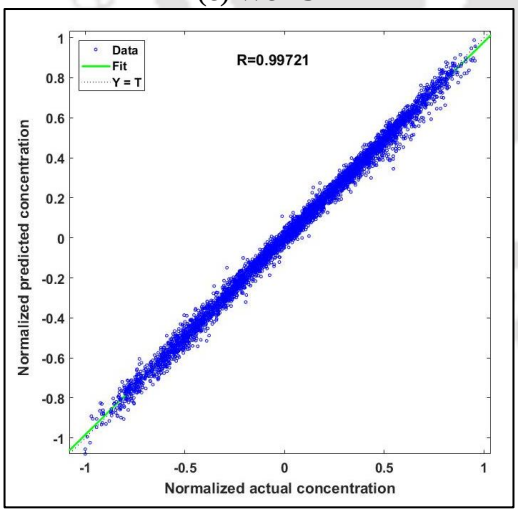
(d) Well 4



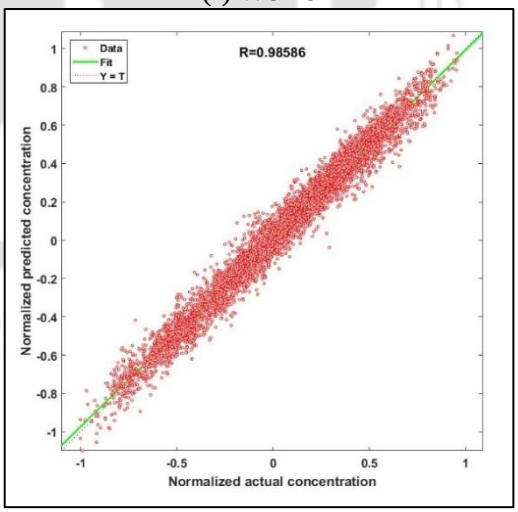
(e) Well 5



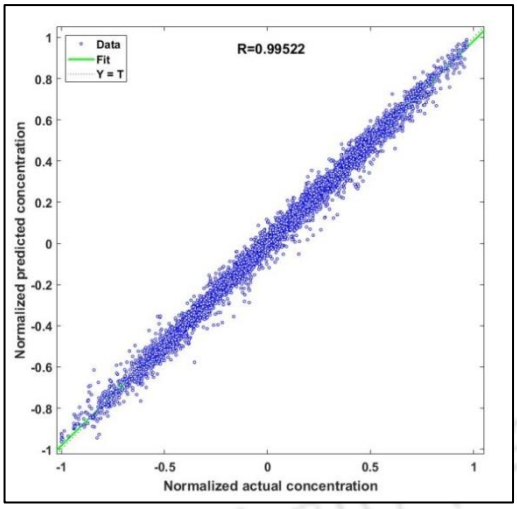
(f) Well 6



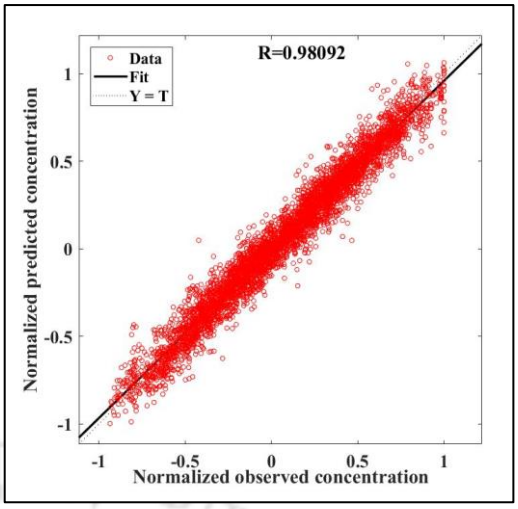
(g) Well 7



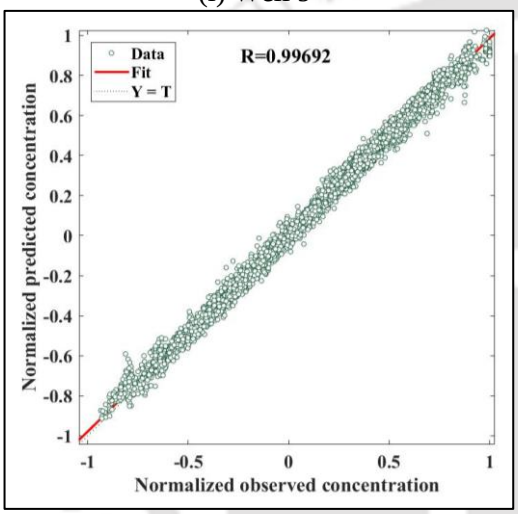
(h) Well 8



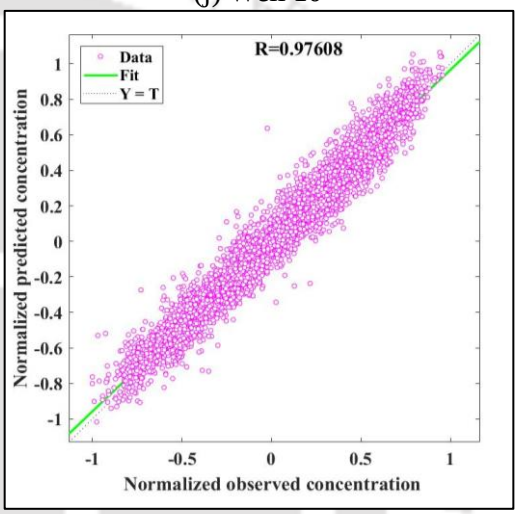
(i) Well 9



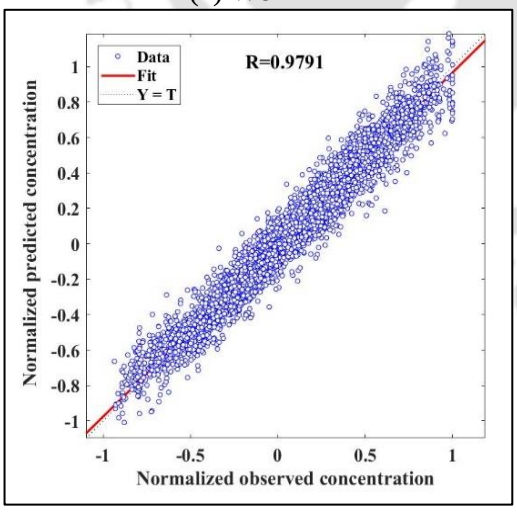
(j) Well 10



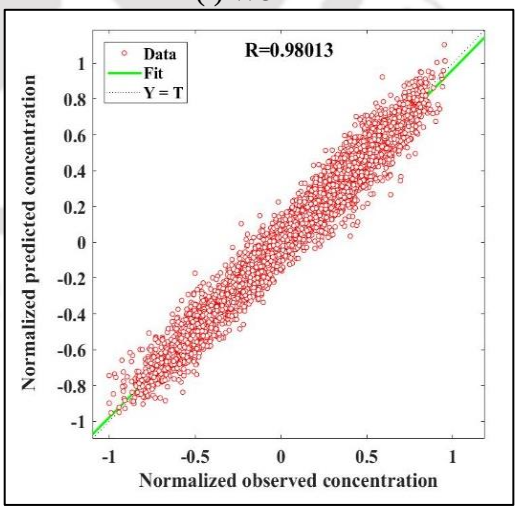
(k) Well 11



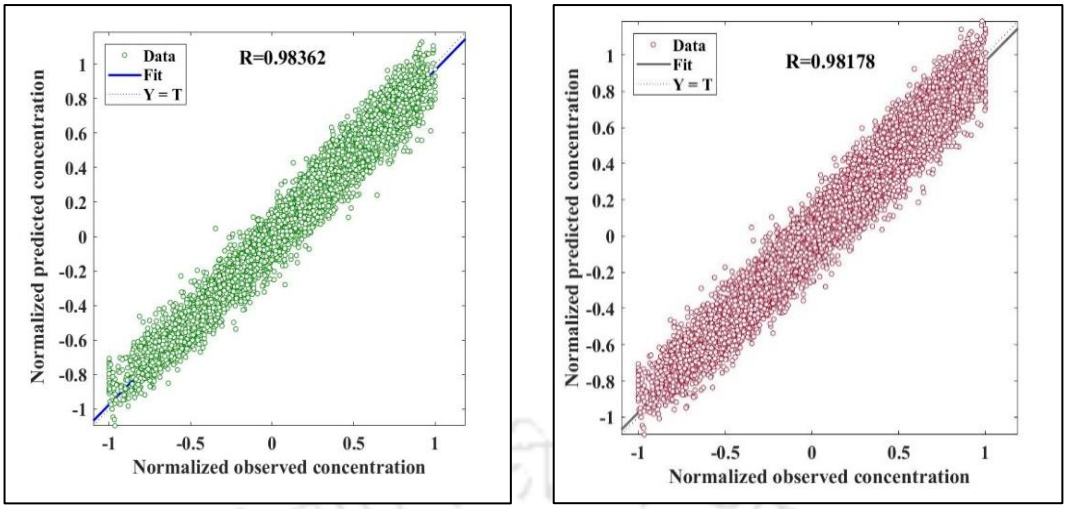
(l) Well 12



(m) Well 13



(n) Well 14



(o) Well 15 (p) Well 16

Figure 6.6: Scatter plots for RNN model developed for 9 observation wells.

Among all observation wells in Figure 6.6, the best and the worst coefficient of correlation (R^2) values are 0.997 and 0.979, respectively. It can be seen that R^2 is very close to 1, which provides a strong correlation between the actual virus concentration and the predicted virus concentration. This observation suggests that the RNN model has successfully captured the patterns and relationships within the dataset generated by the Numerical Transport Simulation Model (NSM). Thus, it can be concluded that the developed RNN model serves as a strong approximate simulator for simulating the coupled groundwater flow and transport model in the proposed study area.

6.3.4 User-defined parameters for SFLA:

This study uses the Shuffled Frog Leaping Algorithm (SFLA) as an optimization algorithm to identify the virus sources. The SFLA algorithm has several parameters that must be considered to optimize the solution effectively. The number of memeplexes was selected as 6, and the number of virtual frogs was selected as the maximum number of variables plus one, which is 8+1, equal to 9. Therefore, 9 virtual frogs for each memeplex were selected, comprising a total of 54 frogs. The maximum step size was taken as 1 unit, which is the maximum distance a frog can jump in one step. The maximum number of evolutions in each memeplex is 8, which is the number of iterations before shuffling the frogs in each memeplex. The step length coefficient is 2, which determines the maximum distance that a frog can jump based on its fitness value. The maximum number of iterations is restricted to 200, which is the maximum number of times the algorithm will iterate before it stops.

6.3.5 Performance of RNN-SFLA-based source identification model (No. of decision variable = 8):

The source identification model is initially developed by coupling the RNN model with the Shuffled Frog Leaping Algorithm (SFLA). The objective is to identify the unknown variables, which include the input source flux (*If*), source locations (*Sl*), and source strength (*Sf*). Here, the source locations are identified not by grid IDs but by their coordinates in the XY plane. Therefore, the number of decision variables in the source identification model is 8, corresponding to the unknown variables. Table 6.2 presents the estimated source locations and fluxes obtained using the RNN-SFLA-based source identification model. The analysis reveals that the RNN-SFLA-based source identification model accurately predicts the source locations as Sl_1 at coordinates (4,5,1) and Sl_2 at (8,7,1).

Table 6.2: Results of Source identification model using RNN-SFLA.

	Sl_1	Sl_2	If_1 (m/h)	If_2 (m/h)	Sf_1 (g/h)	Sf_2 (g/h)	Function value	Time
Lower bound	(1,1)	(1,1)	0.2	0.2	300	200		
Upper bound	(20,20)	(20,20)	0.5	0.5	700	600		232 minutes 26 seconds
Actual	(4,5)	(8,7)	0.35	0.37	500	400		
Predicted	(4,5)	(8,7)	0.388	0.326	532.64	313.23	1.42E-01	
Relative error (%)	---	---`	10.86	11.89	6.52	21.69		

However, when estimating the input source fluxes and source strength, the model provides solutions that are near the optimal values but not exactly equal to them. The estimated virus source strengths are 532.64 g/h and 313.23 g/h, while the estimated input source fluxes are 0.388 m/h and 0.326 m/h. The objective function value is found to be 1.42E-01. This value is the best function value obtained, and further improvements are not observed. Thus, the function tolerance (stopping criteria) compels the simulation to end at 121 iteration steps with 6534 function evaluation. While calculating the relative error, it is observed that the error ranges from 6.52% to 21.69%. This indicates that there is some discrepancy between the predicted values and the actual optimal values. However, it is noted that the relative errors are higher compared to the model developed in Chapter 5, suggesting that further improvements can be made in the accuracy of the estimation. The RNN-SFLA-based source identification model demonstrates computational efficiency, completing 6534 function evaluations in only 232 minutes and 26 seconds. This computational speed is advantageous, as it reduces the time

required for the optimization process. The performance of the RNN-SFLA-based source identification model and the SFLA-NSM (Numerical Transport Simulation Model) is discussed in Section 6.4.

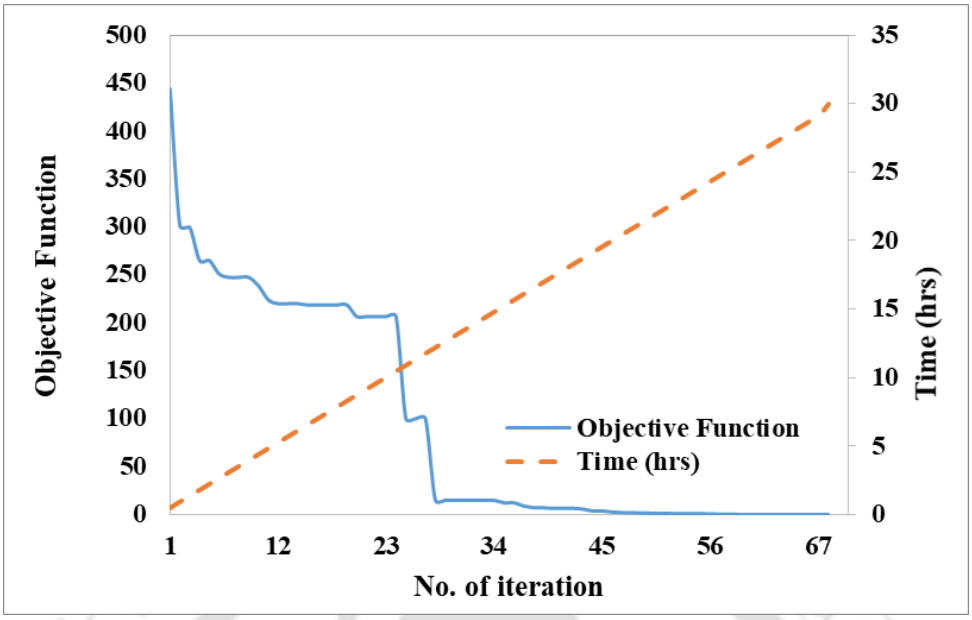
6.4 Comparison of RNN-SFLA and Simulation model-SFLA

Table 6.3 compares the solutions obtained from the RNN-SFLA and NSM-SFLA-based source identification models. It is observed that both models can accurately estimate the source locations. However, when comparing the input source flux and source strength, some percentage of error is present in the solutions obtained by both models.

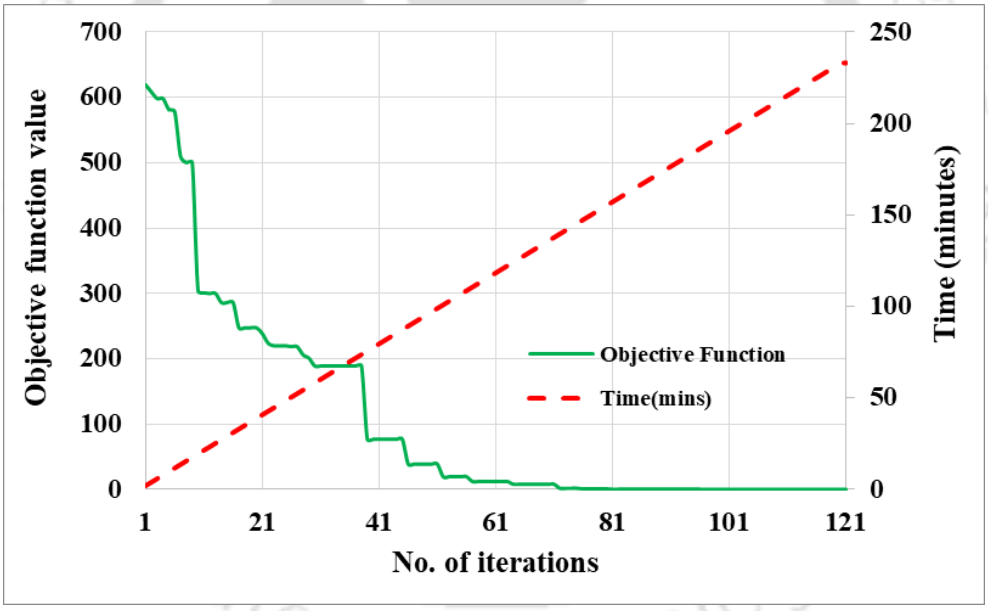
Table 6. 3: Comparison of estimated Source location (Sl), Source strength (Sf), and Input source flux (If) are unknown. Using RNN-SFLA and NSM-SFLA model.

Variable	Actual	Predicted		Relative error (%)	
		RNN-SFLA	NSM-SFLA	RNN-SFLA	NSM-SFLA
Sl_1	(4,5)	(4,5)	(4,5)	---	---
Sl_2	(8,7)	(8,7)	(8,7)	---	---
If_1	0.35	0.388	0.341	10.86	2.571
If_2	0.37	0.326	0.361	11.89	2.432
Sf_1	500	532.64	516.92	6.52	3.384
Sf_2	400	313.23	388.23	21.69	2.94

Estimation of the relative error percentage for both models reveals that the RNN-SFLA-based source identification model provides solutions with an error percentage ranging from 6.52% to 21.69%. On the other hand, the SFLA-NSM model yields solutions with a lower percentage of error, ranging from 2.432% to 3.384%. The lower percentage of error in the SFLA-NSM model is considered acceptable, indicating higher accuracy in estimating the source fluxes (*If*) and source strength (*Sf*). Based on the relative error percentages, it can be concluded that the performance of the SFLA-NSM source identification model is better than that of the RNN-SFLA model in terms of performance efficiency.



(a) Numerical simulation model (NSM)–SFLA



(b) RNN-SFLA

Figure 6. 7: Comparison of objective function value at each iteration with respect to time for (a) NSM-SFLA and (b) RNN-SFLA-based source identification model.

To compare the computational time efficiency, the objective function and computational time required for both models were plotted with respect to the number of iterations, as shown in Figure 6. 7.Figure 6. 7 shows that the RNN-SFLA based source identification model requires a significantly shorter computational time than the SFLA-NSM model. The RNN-SFLA model completes the optimization process in approximately 3.8 hours, while the SFLA-NSM model takes around 29 hours, which is approximately 7 times longer. This model was run in computer

system with specification of processor: Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz 3.41 GHz, 8GB RAM. Analyzing the objective function values, it is noted that the RNN-SFLA model reaches a function value of 1.42E-01 after 121 iterations. In contrast, the SFLA-NSM model achieves an optimal value at 68 iterative steps with a function value of 5.33E-03. Considering the overall investigation, it can be concluded that although the computational time is significantly reduced in the RNN-SFLA-based source identification model, the predicting efficiency of the model has decreased substantially. This indicates that the RNN-SFLA model may not accurately estimate the source strength (*Sf*) and input source fluxes (*If*). To overcome this limitation and strike a balance between computational time and predicting performance efficiency, an improved source identification model is proposed. This model combines the SFLA-based simulation optimization approach with both RNN and NSM. The proposed model seeks to optimize the trade-off between computational efficiency and accuracy, ensuring reliable and efficient source identification results. Further details and analysis of this proposed model are likely discussed in section 6.5.

6.5 RNN- NSM-SFLA-based virus source identification model (Model 2).

From section 6.4, it can be concluded that the RNN-SFLA-based source identification model reach the optimal solution with less computational time, but the predicting efficiency of the model has been reduced substantially. To address this limitation, a new methodology is proposed in this section to develop an improved source identification model. This improved model incorporates the RNN and NSM along with the optimization model, which is the Shuffled Frog Leaping Algorithm (SFLA).

6.5.1 Methodology

In the proposed methodology, the fitness value of the best individual (or frogs) is calculated using the NSM (Numerical Transport Simulation Model), while the fitness of the remaining individuals is calculated using the RNN model. This modification in the source identification model affects the computational time and enables the model to obtain the global optimal solution. The methodology is implemented in the same study area depicted in Figure 6. 3, ensuring consistency and comparability with previous analyses. The objective function used in this modified model is the same as equation (6.1). The parameters used for the RNN and SFLA components remain the same as those considered in Section 6.3. To provide a visual representation of the proposed methodology, a flowchart of the modified source identification model developed using RNN-NSM-SFLA is presented in Figure 6. 8. The flowchart likely

outlines the sequential steps and processes involved in the model, including data input, training and optimization, fitness evaluation, and obtaining the optimal solution.

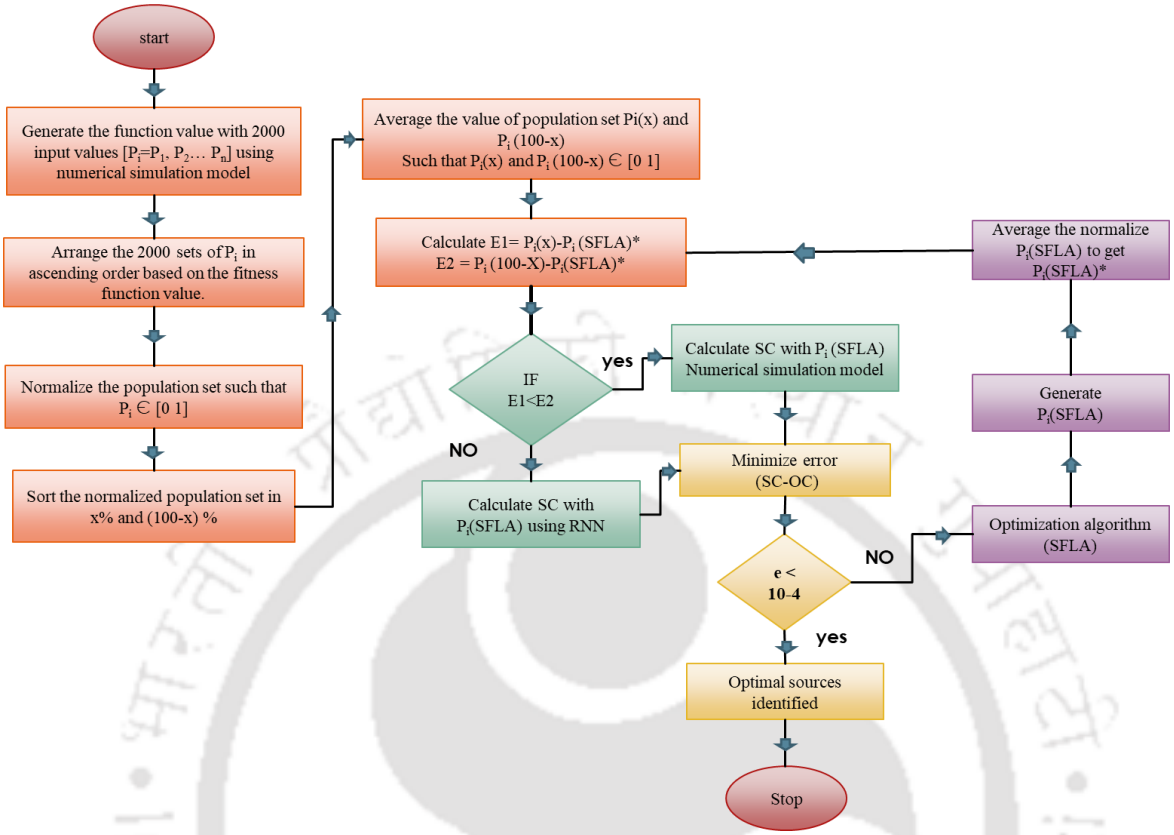


Figure 6. 8: Flowchart of the modified source identification model developed using the SFLA-RNN- Numerical Transport Simulation Model (NSM)

Here is a summary of the steps involved in the method:

1. Generate 2000 sets of input-output data using an NSM.
2. Calculate the fitness values for each set of the population using the generated data.
3. Sort the population set [p_1, p_2….p_n] in ascending order based on their fitness values. The population with lower fitness values (better solutions) is listed first, and those with higher fitness values (poorer solutions) are listed last.
4. Divide the sorted population into two parts: the most effective solutions lies in x% group (where x = 20%, 40%, 60%, or 80%), and the rest in (100 - x)% group.
5. Apply the Shuffled Frog Leaping algorithm to generate a population set and determine whether the solutions lie within the x% or (100 - x)% population group.
6. There is a switch criterion where the population from the SFLA decides whether to move to the RNN model or the NSM to calculate the simulated concentrations.

7. The yellow color in the flow chart represents the objective function and termination criteria. If the termination criteria are satisfied, the optimal solution is obtained. If not, the SFLA is used to generate a new population set (P_i (SFLA)).
8. The population sets (P_i (SFLA)) are normalized to the range [0, 1], as indicated by the purple color in the flow chart. The normalized population sets are used to calculate the Error E_1 and E_2, which influence the switch criteria represented by the green color. This process is repeated until the optimal solution is obtained.

It is to be noted that the source location (*Sl*), input source fluxes (*If*), and source strength (*Sf*) within the 2000 sets of the population are normalized ranging from [0, 1] based on the minimum and maximum values within those sets,

6.5.2 Observation of source identification model using RNN-NSM-SFLA (Model 2):

Table 6. 4 shows the predicted virus source location and fluxes obtained using the SFLA-RNN-NSM method in Scenario 1. The method uses different values of *x*, which determines whether the SFLA-NSM model or the SFLA-RNN model is used for predictions. When *x* is 0%, indicating the SFLA-RNN model, the RNN model is used. When *x* is 100%, indicating the SFLA-NSM model, the NSM is used to calculate the simulated concentrations. The results indicate that the virus source location (*Sl*) was accurately estimated regardless of the *x* value used. However, there were variations observed in the predicted input source fluxes (*If*) and source strength (*Sf*) for different *x* values. The model performed best when *x* was equal to 60%, as it can accurately predict the virus source fluxes and source strength with minimum relative error. Therefore, the SFLA-RNN-NSM method effectively predicts virus source location and fluxes, with the choice of *x* value affecting the accuracy of the results.

Table 6. 4: Predicted source location (*Sl*) and input source fluxes (*If*) and source strength (*Sf*) using SFLA-based RNN-NSM with increased number of *x*%.

Variables	**Actual**	**Predicted (RNN + Simulation model + SFLA) with x%**				
		20%	**40%**	**60%**	**80%**	**100%**
Sl_1	(4,5)	(4,5)	(4,5)	(4,5)	(4,5)	(4,5)
Sl_2	(8,7)	(8,7)	(8,7)	(8,7)	(8,7)	(8,7)
If_1	0.35	0.3061	0.3751	0.370	0.3498	0.353
If_2	0.37	0.462	0.327	0.366	0.379	0.3701
Sf_1	500	489.53	512.64	509. 04	488.38	497.92
Sf_2	400	410.26	378.64	400.23	398.66	399.23

The relative error (%) between the estimated and predicted source location and fluxes is calculated for different values of x: 20%, 40%, 60%, 80%, and 100%. The relative error is then plotted with respect to the x % and is shown in Figure 6. 9. The analysis shows that when the *x* value is equal to 20%, the error percentage is high and falls outside the acceptable range. However, as the value of *x* increases, the percentage of error declines. Specifically, when *x* becomes more than 60%, the model exhibits relative errors ranging from 0.057% to 5.174%. This indicates that the predicting performance of the source identification model using SFLA-RNN-NSM is superior to that of the SFLA-RNN model. Additionally, the objective function versus computational time is plotted for different *x* percentages in Figure 6.10 to assess computational time efficiency.

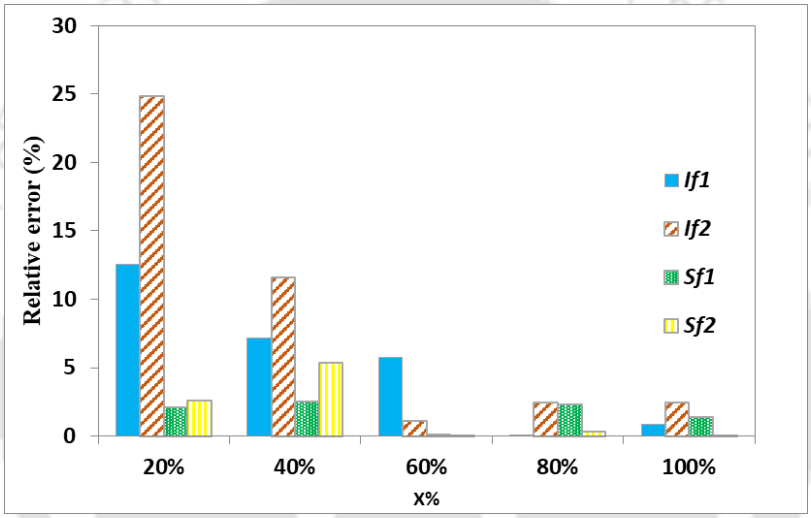


Figure 6. 9: Relative error obtained between the actual and predicted sources at different X%.

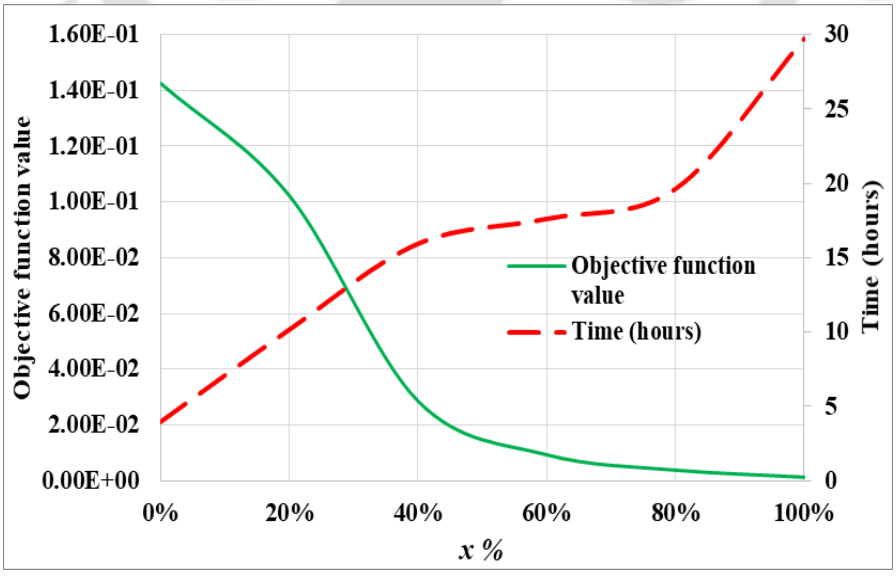


Figure 6. 10: Variation of objective function value and computational time with x%.

From Figure 6. 10, the evaluation shows that as the *x*% value increases, the objective function decreases while the computational time increases. This indicates a trade-off between the accuracy of the solution and the computational efficiency. In the case of x = 60%, which provided accurate predictions for virus source fluxes and input velocity fluxes, the computational time was monitored and found to be approximately 17 hours. Although better solutions were obtained at x = 80% and 100%, the computational time required was higher compared to the model developed using x = 60%. Figure 6. 11 represents the total number of function evaluations and the model used at each evaluation.

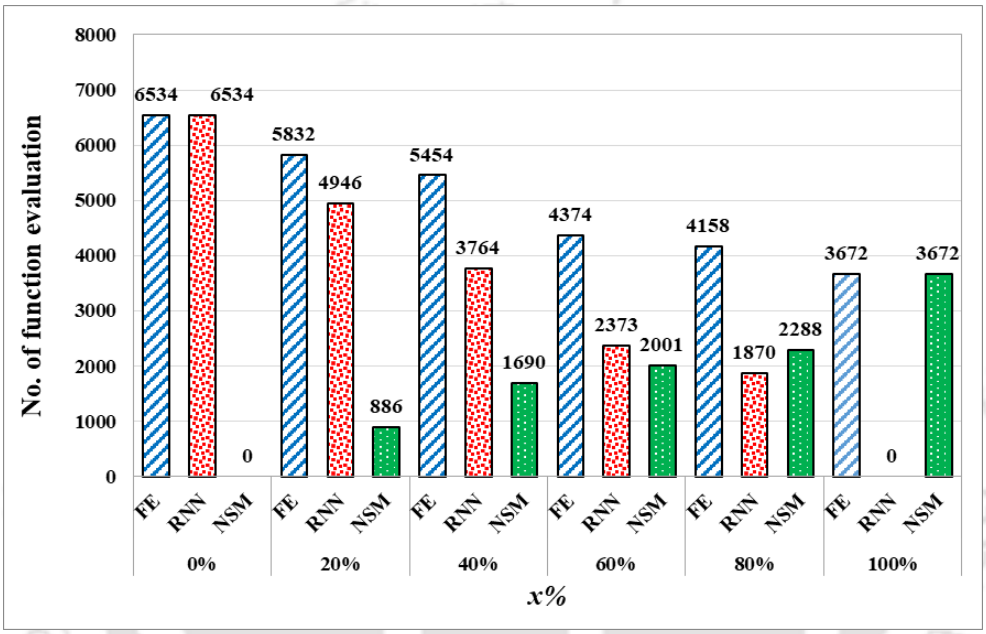


Figure 6. 11: Total function evaluation (FE) count with respect to RNN and Numerical Transport Simulation Model (NSM) along with the increase in x%.

Figure 6. 11 provides insights into the evaluation process and the models employed during the optimization. Based on the information provided, the following observations can be made:

1. When x = 0%, the number of function evaluations (FE) is 6534, indicating that the source identification model uses the RNN-SFLA model for calculating simulated concentrations. However, this results in providing a solution with maximum relative error compared to other values of x.
2. When *x* = 100%, the number of function evaluations is 3672, and the model consistently uses the SFLA-NSM approach for every function evaluation. This leads to efficient prediction of the virus sources, indicating that the SFLA-NSM model is effective in this scenario.
3. There is a decreasing trend in the number of function evaluations that utilize the RNN model as *x* increases. Conversely, there is an increasing trend in the use of the SFLA-

NSM model as x increases. At x = 60%, the total number of function evaluations is 4374. Out of these, 2373 function evaluations are performed by the RNN-SFLA model, and the SFLA-NSM model conducts the remaining 2001 function evaluations. This implies that the RNN-SFLA model performs more than 54% of the function evaluations.

4. Using the RNN-SFLA model for more than 54% of the function evaluations at x = 60% results in reduced computational time. At the same time, the function evaluation using SFLA-NSM provides the best solutions. Based on these observations, it appears that the RNN-NSM-SFLA approach, particularly with x = 60%, provides a good balance between computational efficiency and predicting performance for the source identification task in this study.

6.6 Conclusion:

In summary, this study initially developed a source identification model using RNN-SFLA for a known scenario with a two-dimensional search space but unsatisfactory predictive performance. To address this, a novel simulation-optimization methodology combining RNN, NSM, and SFLA was introduced, effectively improving computational efficiency and prediction accuracy. The modified model dynamically selects between RNN and NSM based on fitness values, enhancing the identification of potential virus sources in groundwater aquifers.

- In conclusion, the utilization of RNN-NSM-SFLA model, demonstrates notable advantages in terms of computational time efficiency and accuracy for source identification.
- When considering a significant proportion of function evaluations, approximately 60%, the RNN-SFLA model performs most of these evaluations, thereby reducing computational time. Complementarily, the NSM-SFLA model handles the remaining function evaluations, contributing to determining the global optimal solution.
- This approach contrasts with the limitations of previous models like the GMS-GA-ANN based model (Borah & Bhattacharjya, 2016) , which could only identify source strength without estimating source location.

- The versatility of the RNN-NSM-SFLA model can be extended to three-dimensional space, facilitating the search for source locations by incorporating three variables corresponding to each coordinate axis (x, y, z).
- Therefore, the RNN-NSM-SFLA based source identification model emerges as a promising advancement in efficiently and accurately locating sources while considering their fluxes in environmental or similar scenarios.

CHAPTER 7
SUMMARY, CONCLUSIONS AND FUTUREWORK.

7.1 Summary

Groundwater aquifers are essential drinking water sources but can also be vulnerable to virus contamination. Virus contamination in groundwater can pose a significant risk to human health, mainly where groundwater is used for drinking water. Identifying the sources of viruses in groundwater can help determine the appropriate measures needed to prevent or mitigate contamination, protect public health, and maintain the quality of groundwater resources. In this study, an effective source identification model for viruses in an unconfined groundwater aquifer is developed considering sources in both the unsaturated and saturated zones. The numerical simulation model is needed to obtain the simulated virus concentration. There are several models available for simulation of contaminant transport in an aquifer. Some of the widely used models are HYDRUS-3D, MT3DMS, RT3DMS, etc. These models mainly rely on the constant detachment model. However, this approach may not always be accurate since the detachment-attachment rates of viruses are known to vary based on the saturation level of the medium. To overcome these limitations, a virus transport model that incorporates variations in input parameters based on the degree of saturation and air-water interfaces in unsaturated conditions is considered. The developed model is validated using experimental data to assess its accuracy in predicting virus transport scenarios.

Once the virus transport model is developed, it is linked with an optimization model for source identification in the aquifer. The objective function is to minimize the error between observed and simulated virus concentrations. To identify the virus source location and source strength efficiently, Shuffled Frog Leaping Algorithm is used as an optimization algorithm. The source identification model is developed for two scenarios: one considering the source location in the unsaturated zone and another considering the source location in both the saturated and unsaturated zones. This linked simulation optimization model can find the accurate virus source locations and source fluxes but requires significant computational time. As such, an improved source identification model is developed to overcome the computational limitation by combining a Recurrent Neural Network (RNN) with the numerical simulation model and the Shuffled Frog Leaping Algorithm. This model is efficient in computation time and predicting performance, making it a valuable tool for identifying virus sources in groundwater aquifers. In both flow and virus transport models, several key parameters describe

the behavior and characteristics of the flow and virus in the environment. However, some of these parameters are unavailable on a field or laboratory scale, requiring parameter estimation. As such effective parameter estimation model is developed using a metaheuristic algorithm to estimate flow and virus transport parameters with minimum relative error. After parameter estimation, global sensitivity analysis is performed to assess the sensitivity of the model output to each estimated parameter. This helps researchers and practitioners to identify the most influential parameters and prioritize their efforts to improve accuracy. The conclusion drawn from each of the objectives is discussed in section 7.2.

7.2 Conclusions:

This thesis work consists of four objectives. The conclusion drawn from each of the objectives is discussed below:

Objective 1:

- The three-dimensional virus transport model is developed using block-centered finite difference approximation considering both unsaturated and saturated zones. This model produces good correlation values with the available experimental data.
- The developed models can handle the variations in input parameters based on the degree of saturation, which is an advantage over the existing models, such as MT3DMS and HYDRUS.
- The developed virus transport model can be used for parameter estimation and identifying virus sources in a groundwater aquifer.

Objective 2:

- A parameter estimation model is developed to effectively estimate both the unsaturated and saturated flow parameters using ANN-SFLA.
- The model is efficient both in computational time and predicting performance. This model uses a metaheuristic optimization algorithm that overcomes the problem of non-unique solutions.
- Virus transport models have traditionally relied on the constant detachment model to estimate transport parameters. The parameter estimation model is developed by considering variable attachment detachment rates for two saturation levels.

- Flower Pollination Algorithm (FPA) based Simulation-optimization model effectively estimates virus transport parameters, though the model has multiple non-unique optimal solutions.

Objective 3:

- Using the Shuffled Frog Leaping Algorithm as an optimization algorithm, the source identification model effectively determines virus source locations and source fluxes.
- The model can identify the source location, source strength, and source flux in two-dimensional and three-dimensional spaces.
- The model available till now can estimate virus sources in the saturated zone only, but this model can effectively estimate virus source location in a saturated zone and in an unsaturated zone.
- However, this model requires high computational time as the numerical simulation model is called multiple times by the optimization model for function evaluation at each iteration.

Objective 4:

- An improved source identification model is developed by combining RNN with the numerical simulation model (NSM) and the Shuffled Frog Leaping Algorithm (SFLA), which addresses the computational time limitation of the previous model while maintaining high predicting performance.
- The RNN-NSM-SFLA based source identification model offers advantages over model available in literature by reducing the function evaluations and at the same time estimating the source locations along with the source fluxes.
- The developed model can be extended to identify the virus pollution sources in three-dimensional space.

7.3 Research Contribution:

This research work aims to develop a contaminant source identification model in an unconfined aquifer considering unsaturated- saturated zone. The major contribution from this research is listed as follows:

- The research work provided a 3-D virus transport model that can study the movement of viruses from unsaturated zone to the groundwater aquifer. As the virus fate and

transport model depends upon the degree of saturation, the model developed provides an innovative approach that accounts for the influence of saturation levels on virus movement, providing a more accurate representation of virus fate and transport in groundwater systems.

- While developing any groundwater flow and transport model, there are different parameter which needs to be incorporated; among them, a few parameters cannot estimated in field scale or laboratory study. As such, an effective parameter estimation model is develop with a novel technique for both the flow and transport model which can estimate the unsaturated and saturated parameters simultaneously.
- Groundwater contamination primarily originates from sources within the unsaturated zone in real-world scenarios. However, previous research has been deficient in pinpointing the precise location of these sources within the unsaturated zone. Therefore, this study significantly contributes to groundwater management by introducing a source identification model capable of accurately locating contamination sources within the unsaturated zone.
- In this study, a small area is hypothetically considered for model development. However, when applying this model in practical fields, one can expand the search area by increasing the number and size of grids. Once a location is detected, that area is earmarked for the future refinement of the source identification model, involving adjustments to grid dimensions and coverage area. This approach aims to enhance the precision of detecting the source location within that designated area.

7.4 Limitation of the Work:

A few limitations and assumptions were made to fulfill the above objectives. By addressing these limitations and assumptions, future research can advance the accuracy and applicability of virus transport models in groundwater. This can lead to more robust strategies for preventing contamination, protecting public health, and managing groundwater resources effectively.

- ✓ The virus transport model developed in this study is limited to a homogeneous soil profile. Further model modification is required to deal with the soil heterogeneity.
- ✓ Due to a lack of experimental data, the three-dimensional virus transport model is limited by considering variations in transport parameters for two saturation levels. However, if parameter variation for each degree of saturation level is available, then the model will be more realistic.

- ✓ The improved source identification model using RNN-NSM-SFLA is limited to search the sources in two-dimensional search space only. Future work can continue by identifying the source location using the improved source identification model in three-dimensional space.

REFERENCES

Abdelgawad, H. A. A. (2009). Parameter estimation of pumping test data using genetic algorithm groundwater resources management and contaminant transport in coastal aquifers: application on quaternary aquifer east Nile delta, egypt View project Confined seepage using Differential Quadrature Element Method View project. https://www.researchgate.net/publication/330824993

Afzaal, H., Farooque, A. A., Abbas, F., Acharya, B., & Esau, T. (2020). Groundwater estimation from major physical hydrology components using artificial neural networks and deep learning. *Water (Switzerland)*, *12*(1). https://doi.org/10.3390/w12010005

Anders, R., & Chrysikopoulos, C. V. (2009). Transport of viruses through saturated and unsaturated columns packed with sand. *Transport in Porous Media*, *76*(1). https://doi.org/10.1007/s11242-008-9239-3

Aral, M. M., & Guan, J. (1996). Genetic Algorithms in Search of Groundwater Pollution Sources. In *Advances in Groundwater Pollution Control and Remediation*. https://doi.org/10.1007/978-94-009-0205-3_17

Aral, M. M., Guan, J., & Maslia, M. L. (2001). Identification of Contaminant Source Location and Release History in Aquifers. *Journal of Hydrologic Engineering*, *6*(3). https://doi.org/10.1061/(asce)1084-0699(2001)6:3(225)

Bales, R. C., Hinkle, S. R., Kroeger, T. W., Stocking, K., & Gerba, C. P. (1991). Bacteriophage Adsorption during Transport through Porous Media: Chemical Perturbations and Reversibility. *Environmental Science and Technology*, *25*(12), 2088–2095. https://doi.org/10.1021/es00024a016

Bales, R. C., Li, S., Yeh, T. C. J., Lenczewski, M. E., & Gerba, C. P. (1997). Bacteriophage and microsphere transport in saturated porous media: Forced-gradient experiment at Borden, Ontario. *Water Resources Research*, *33*(4). https://doi.org/10.1029/97WR00025

Bedekar, V., Morway, E., Langevin, C., & Tonkin, M. (2016). MT3D-USGS version 1: A U.S. Geological Survey release of MT3DMS updated with new and expanded transport capabilities for use with MODFLOW. *Groundwater Resources Program*.

Bhattacharjee, S., Ryan, J. N., & Elimelech, M. (2002). Virus transport in physically and geochemically heterogeneous subsurface porous media. *Journal of Contaminant Hydrology*, *57*(3–4), 161–187. https://doi.org/10.1016/S0169-7722(02)00007-4

Bhattacharjya, R. K., & Datta, B. (2005). Optimal management of coastal aquifers using linked simulation optimization approach. *Water Resources Management*, *19*(3), 295–320. https://doi.org/10.1007/s11269-005-3180-9

Bhattacharjya, R. K., & Datta, B. (2009). ANN-GA-Based Model for Multiple Objective Management of Coastal Aquifers. *Journal of Water Resources Planning and Management*, *135*(5), 314–322. https://doi.org/10.1061/(asce)0733-9496(2009)135:5(314)

Bhattacharjya, R. K., Datta, B., & Satish, M. G. (2007). Artificial Neural Networks Approximation of Density Dependent Saltwater Intrusion Process in Coastal Aquifers. *Journal of Hydrologic Engineering*, *12*(3). https://doi.org/10.1061/(asce)1084-0699(2007)12:3(273)

Bhattacharjya, R. K., Srivastava, A., & Satish, M. G. (2015). Hybrid-Optimization Approach for Estimating Parameters of a Virus Transport Process in Aquifer. *Journal of Hazardous, Toxic, and Radioactive Waste*, *19*(2), 04014025. https://doi.org/10.1061/(asce)hz.2153-5515.0000195

Borah, T., & Bhattacharjya, R. K. (2013). Solution of source identification problem by using GMS and MATLAB. *ISH Journal of Hydraulic Engineering*, *19*(3), 297–304. https://doi.org/10.1080/09715010.2013.808515

Borah, T., & Bhattacharjya, R. K. (2015). Development of Unknown Pollution Source Identification Models Using GMS ANN–Based Simulation Optimization Methodology. *Journal of Hazardous, Toxic, and Radioactive Waste*, *19*(3), 04014034. https://doi.org/10.1061/(asce)hz.2153-5515.0000242

Borah, T., & Bhattacharjya, R. K. (2016). Development of an Improved Pollution Source Identification Model Using Numerical and ANN Based Simulation-Optimization Model. *Water Resources Management*, *30*(14). https://doi.org/10.1007/s11269-016-1476-6

Bouhlila, R., & Hariga, N. T. (2022). Identification of aquifer pollution's point sources with the reciprocity principle. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-13795-w

Brooks, R., & Corey, A. (1964). Hydraulic properties of porous media. *Hydrology Papers, Colorado State University*, *3*(March).

Burnett, R. D., & Frind, E. O. (1987). Simulation of contaminant transport in three dimensions: 2. Dimensionality effects. *Water Resources Research*, *23*(4). https://doi.org/10.1029/WR023i004p00695

Cavazzuti, M. (2013). Optimization methods: From theory to design scientific and technological aspects in mechanics. In *Optimization Methods: From Theory to Design Scientific and Technological Aspects in Mechanics*. https://doi.org/10.1007/978-3-642-31187-1

Celia, M. A., Bouloutas, E. T., & Zarba, R. L. (1990). A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resources Research*, *26*(7), 1483–1496. https://doi.org/10.1029/WR026i007p01483

Chen, H., Zhang, W., Nie, N., & Guo, Y. (2019). Long-term groundwater storage variations estimated in the Songhua River Basin by using GRACE products, land surface models, and in-situ observations. *Science of the Total Environment*, *649*. https://doi.org/10.1016/j.scitotenv.2018.08.352

Chrysikopoulos, C. V., & Sim, Y. (1996). One-dimensional virus transport in homogeneous porous media with time- dependent distribution coefficient. *Journal of Hydrology*, *185*(1–4), 199–219. https://doi.org/10.1016/0022-1694(95)02990-7

Chrysikopoulos, C. V., & Vogler, E. T. (2004). Estimation of time dependent virus inactivation rates by geostatistical and resampling techniques: Application to virus transport in porous media. *Stochastic Environmental Research and Risk Assessment*, *18*(2), 67–78. https://doi.org/10.1007/s00477-003-0130-z

Chu, Y., Jin, Y., Flury, M., & Yates, M. V. (2001). Mechanisms of virus removal during transport in unsaturated porous media. *Water Resources Research*, *37*(2), 253–263. https://doi.org/10.1029/2000WR900308

Clement, T. P., Wise, W. R., & Molz, F. J. (1994). A physically based, two-dimensional, finite-difference algorithm for modeling variably saturated flow. *Journal of Hydrology*, *161*(1–4). https://doi.org/10.1016/0022-1694(94)90121-X

Craun, G. F. (1992). Waterborne disease outbreaks in the United States of America: Causes and prevention. *World Health Statistics Quarterly*, *45*(2–3), 192–199.

Dane, J. H., & Hruska, S. (1983). In-Situ Determination of Soil Hydraulic Properties during Drainage. *Soil Science Society of America Journal*, *47*(4). https://doi.org/10.2136/sssaj1983.03615995004700040001x

Datta, B., Prakash, O., & Sreekanth, J. (2014). Application of genetic programming models incorporated in optimization models for contaminated groundwater systems management. *Advances in Intelligent Systems and Computing*, *288*. https://doi.org/10.1007/978-3-319-07494-8_13

Dogan, A., & Motz, L. H. (2005). Saturated-Unsaturated 3D Groundwater Model. II: Verification and Application. *Journal of Hydrologic Engineering*, *10*(6). https://doi.org/10.1061/(asce)1084-0699(2005)10:6(505)

Dogan, A., Motz, L. H., & Asce, M. (2005). *I : Development*. *10*(December), 492–504. https://doi.org/10.1061/(ASCE)1084-0699(2005)10

Eching, S. O., & Hopmans, J. W. (1993). Optimization of Hydraulic Functions from Transient Outflow and Soil Water Pressure Data. *Soil Science Society of America Journal*, *57*(5). https://doi.org/10.2136/sssaj1993.03615995005700050001x

Eching, S. O., Hopmans, J. W., & Wendroth, O. (1994). Unsaturated Hydraulic Conductivity from Transient Multistep Outflow and Soil Water Pressure Data. *Soil Science Society of America Journal*, *58*(3). https://doi.org/10.2136/sssaj1994.03615995005800030008x

El Amri, A., M'nassri, S., Nasri, N., Nsir, H., & Majdoub, R. (2022). Nitrate concentration analysis and prediction in a shallow aquifer in central-eastern Tunisia using artificial neural network and time series modelling. *Environmental Science and Pollution Research*, *29*(28). https://doi.org/10.1007/s11356-021-18174-y

Fredriksson, L. (n.d.). *A Brief Survey of Lévy Walks with applications to probe diffusion*. www.kau.se

Gassilloud B, Gantzer C (2017). Adhesion-aggregation and inactivation of Poliovirus 1 in groundwater stored in a hydrophobic container. *Appl Environ Microbiol.71(2):912–20. pmid:15691948*

Gaur, S., Ch, S., Graillot, D., Chahar, B. R., & Kumar, D. N. (2013). Application of Artificial Neural Networks and Particle Swarm Optimization for the Management of Groundwater Resources. *Water Resources Management*, *27*(3). https://doi.org/10.1007/s11269-012-0226-7

Gerba, C. P. (1984). Applied and Theoretical Aspects of Virus Adsorption to Surfaces. *Advances in Applied Microbiology*, *30*(C). https://doi.org/10.1016/S0065-2164(08)70054-6

Gerba CP, Choi CY (2006). Role of irrigation water in crop contamination by viruses. *In: Viruses in Foods.* p. 7.

Goh, A. T. C., Zhang, R. H., Wang, W., Wang, L., Liu, H. L., & Zhang, W. G. (2020). Numerical study of the effects of groundwater drawdown on ground settlement for excavation in residual soils. *Acta Geotechnica*, *15*(5). https://doi.org/10.1007/s11440-019-00843-5

Gorelick, S. M., Evans, B., & Remson, I. (1983). Identifying sources of groundwater pollution: An optimization approach. *Water Resources Research*, *19*(3). https://doi.org/10.1029/WR019i003p00779

Gotkowitz MB, Bradbury KR, Borchardt MA, Zhu J, Spencer SK (2016). Effects of Climate and Sewer Condition on Virus Transport to Groundwater. *Environ Sci Technol. 50(16):8497–504. pmid:27434550*

Gribb, M. M. (1996). Parameter estimation for determining hydraulic properties of a fine sand from transient flow measurements. *Water Resources Research*, *32*(7). https://doi.org/10.1029/96WR00894

Guan, J., & Aral, M. M. (1999). Optimal remediation with well locations and pumping rates selected as continuous decision variables. *Journal of Hydrology*, *221*(1–2). https://doi.org/10.1016/S0022-1694(99)00079-7

Guneshwor, L., Eldho, T. I., & Vinod Kumar, A. (2018). Identification of Groundwater Contamination Sources Using Meshfree RPCM Simulation and Particle Swarm Optimization. *Water Resources Management*, *32*(4). https://doi.org/10.1007/s11269-017-1885-1

Gurarslan, G., & Karahan, H. (2015). Solving inverse problems of groundwater-pollution-source identification using a differential evolution algorithm. *Hydrogeology Journal*, *23*(6). https://doi.org/10.1007/s10040-015-1256-z

Häfliger, D., Hübner, P., & Lüthy, J. (2000). Outbreak of viral gastroenteritis due to sewage-contaminated drinking water. *International Journal of Food Microbiology*, *54*(1–2). https://doi.org/10.1016/S0168-1605(99)00176-2

Huang, C. L., & Dun, J. F. (2008). A distributed PSO-SVM hybrid system with feature selection and parameter optimization. *Applied Soft Computing Journal*, *8*(4). https://doi.org/10.1016/j.asoc.2007.10.007

Hunt RJ, Johnson WP (2017). Transport de pathogènes dans les systèmes aquifères: contrastes avec le transports traditionnel de solutés. Hydrogeol J. ;25(4):921–30.

Huo, Z. L., Kang, S. Z., Cen, S. J., & Ma, Y. (2007). Simulation of effects of agricultural activities on groundwater level by combining FEFLOW and GIS. *New Zealand Journal of Agricultural Research*, *50*(5). https://doi.org/10.1080/00288230709510358

Hurst, C. J., Gerba, C. P., & Cech, I. (1980). Effects of environmental variables and soil characteristics on virus survival in soil. *Applied and Environmental Microbiology*, *40*(6), 1067–1079. https://doi.org/10.1128/aem.40.6.1067-1079.1980

Hyun, Y., & Lee, K. K. (1998). Model identification criteria for inverse estimation of hydraulic parameters. *Ground Water*, *36*(2). https://doi.org/10.1111/j.1745-6584.1998.tb01088.x

Inoue, M., Šimůnek, J., Shiozawa, S., & Hopmans, J. W. (2000). Simultaneous estimation of soil hydraulic and solute transport parameters from transient infiltration experiments. *Advances in Water Resources*, *23*(7), 677–688. https://doi.org/10.1016/S0309-1708(00)00011-7

Jacques, D., Šimůnek, J., Timmerman, A., & Feyen, J. (2002). Calibration of Richards' and convection-dispersion equations to field-scale water flow and solute transport under rainfall conditions. *Journal of Hydrology*, *259*(1–4), 15–31. https://doi.org/10.1016/S0022-1694(01)00591-1

Jiang, S., Zhang, Y., Wang, P., & Zheng, M. (2013). An almost-parameter-free harmony search algorithm for groundwater pollution source identification. *Water Science and Technology*, *68*(11). https://doi.org/10.2166/wst.2013.499

Jiang, X., Lu, W., Hou, Z., Zhao, H., & Na, J. (2015). Ensemble of surrogates-based optimization for identifying an optimal surfactant-enhanced aquifer remediation strategy at heterogeneous DNAPL-contaminated sites. *Computers and Geosciences*, *84*. https://doi.org/10.1016/j.cageo.2015.08.003

Jin, Y., Chu, Y., & Li, Y. (2000). Virus removal and transport in saturated and unsaturated sand columns. *Journal of Contaminant Hydrology*, *43*(2), 111–128. https://doi.org/10.1016/S0169-7722(00)00084-X

Jin, Y., & Flury, M. (2002). Fate and transport of viruses in porous media. In *Advances in Agronomy* (Vol. 77). https://doi.org/10.1016/s0065-2113(02)77013-2

Joshi, N., Ojha, C. S. P., Sharma, P. K., & Surampalli, R. Y. (2013). Parameter identification of virus transport in porous media using equilibrium and non-equilibrium models. *Journal of Environmental Chemical Engineering*, *1*(4), 1099–1107. https://doi.org/10.1016/j.jece.2013.08.023

Keswick, B. H., & Gerba, C. P. (1980). Viruses in groundwater. *Environmental Science and Technology*, *14*(11), 1290–1297. https://doi.org/10.1021/es60171a602

Kodešová, R., Gribb, M. M., & Šimůnek, J. (1998). Estimating soil hydraulic properties from transient cone permeameter data. *Soil Science*, *163*(6), 436–453. https://doi.org/10.1097/00010694-199806000-00002

Kodešová, R., Ordway, S. E., Gribb, M. M., & Šimůnek, J. (1999). Estimation of soil hydraulic properties with the cone permeameter: Field studies. *Soil Science*, *164*(8). https://doi.org/10.1097/00010694-199908000-00001

Kool, J. B., & Parker, J. C. (1987). Development and evaluation of closed-form expressions for hysteretic soil hydraulic properties. *Water Resources Research*, *23*(1). https://doi.org/10.1029/WR023i001p00105

Kool, J. B., Parker, J. C., & van Genuchten, M. T. (1987). Parameter estimation for unsaturated flow and transport models - A review. *Journal of Hydrology*, *91*(3–4). https://doi.org/10.1016/0022-1694(87)90207-1

Koonin, E. V., Senkevich, T. G., & Dolja, V. V. (2006). The ancient virus world and evolution of cells. *Biology Direct*, *1*. https://doi.org/10.1186/1745-6150-1-29

Kourakos, G., & Mantoglou, A. (2011). Simulation and Multi-Objective Management of Coastal Aquifers in Semi-Arid Regions. *Water Resources Management*, *25*(4). https://doi.org/10.1007/s11269-010-9677-x

Kourakos, G., & Mantoglou, A. (2013). Development of a multi-objective optimization algorithm using surrogate models for coastal aquifer management. *Journal of Hydrology*, *479*. https://doi.org/10.1016/j.jhydrol.2012.10.050

Lance, J. C., & Gerba, C. P. (1984). Virus movement in soil during saturated and unsaturated flow. *Applied and Environmental Microbiology*, *47*(2), 335–337. https://doi.org/10.1128/aem.47.2.335-337.1984

Lappala, E. G., Healy, R. W., & Weeks, E. P. (1987). Documentation of Computer Program VS2D to Solve the Equations of Fluid Flow in Variably Saturated Porous Media. *Usgs*.

Leichombam, S., & Bhattacharjya, R. K. (2016). Identification of Unknown Groundwater Pollution Sources and Determination of Optimal Well Locations Using ANN-GA Based Simulation-Optimization Model. *Journal of Water Resource and Protection*, *08*(03). https://doi.org/10.4236/jwarp.2016.83034

Leichombam, S., & Bhattacharjya, R. K. (2019). New Hybrid Optimization Methodology to Identify Pollution Sources Considering the Source Locations and Source Flux as Unknown. *Journal of Hazardous, Toxic, and Radioactive Waste*, *23*(1). https://doi.org/10.1061/(asce)hz.2153-5515.0000431

Lin, H. J., Richards, D. R., Talbot, C. A., Station, W. E., Yeh, G., Cheng, J., Cheng, H., & Jones, N. L. (1997). FEMWATER : A Three-Dimensional Finite Element Computer Model for Simulating Density-Dependent Flow and Transport in Variably Saturated Media by. *Transport*, *July*.

Liu, H., Li, J., Cao, H., Xie, X., & Wang, Y. (2022). Prediction modeling of geogenic iodine contaminated groundwater throughout China. *Journal of Environmental Management*, *303*. https://doi.org/10.1016/j.jenvman.2021.114249

Loveland, J. P., Ryan, J. N., Amy, G. L., & Harvey, R. W. (1996). The reversibility of virus attachment to mineral surfaces. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, *107*. https://doi.org/10.1016/0927-7757(95)03373-4

Mack, W. N., Lu, Y. S., & Coohon, D. B. (1972). Isolation of poliomyelitis virus from a contaminated well. *Health Services Reports*, *87*(3). https://doi.org/10.2307/4594488

Mahinthakumar, G. (Kumar), & Sayeed, M. (2005). Hybrid Genetic Algorithm—Local Search Methods for Solving Groundwater Source Identification Inverse Problems. *Journal of Water Resources Planning and Management*, *131*(1). https://doi.org/10.1061/(asce)0733-9496(2005)131:1(45)

Maseri, S. Z., Hadi, M. S., Jamali, A., Yatim, H. M., Ab Talib, M. H., & Darus, I. Z. M. (2019, October). A Single objective flower pollination algorithm for modeling the horizontal flexible plate system. *In 2019 2nd International Conference on Applied Engineering (ICAE) (pp. 1-6). IEEE.*

Mathew, T. V. (n.d.). *Genetic Algorithm*.

Matiatos, I., Varouchakis, E. A., & Papadopoulou, M. P. (2019). Performance Evaluation of Multiple Groundwater Flow and Nitrate Mass Transport Numerical Models. *Environmental Modeling and Assessment*, *24*(6). https://doi.org/10.1007/s10666-019-9653-7

McKinney, D. C., & Lin, M. -D. (1994). Genetic algorithm solution of groundwater management models. *Water Resources Research*, *30*(6), 1897–1906. https://doi.org/10.1029/94WR00554

McLaughlin, D., & Townley, L. R. (1996). A reassessment of the groundwater inverse problem. In *Water Resources Research* (Vol. 32, Issue 5). https://doi.org/10.1029/96WR00160

Mohanty, S., Jha, M. K., Kumar, A., & Panda, D. K. (2013). Comparative evaluation of numerical model and artificial neural network for simulating groundwater flow in Kathajodi-Surua Inter-basin of Odisha, India. *Journal of Hydrology*, *495*. https://doi.org/10.1016/j.jhydrol.2013.04.041

Ngamsritrakul, T., Ruangrassamee, P., & Putthividhya, A. (2015). Backward probabilistic model for contaminant source identification in Thailand. *World Environmental and Water Resources Congress 2015: Floods, Droughts, and Ecosystems - Proceedings of the 2015 World Environmental and Water Resources Congress*. https://doi.org/10.1061/9780784479162.043

Noori, R., Khakpour, A., Omidvar, B., & Farokhnia, A. (2010). Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic. *Expert Systems with Applications*, *37*(8). https://doi.org/10.1016/j.eswa.2010.02.020

O'Connor, D. (2002). Part One : A Summary Report of the Walkerton Inquiry : The Events of May 2000 and Related Issues. In *Training* (Issue May).

O'Connor, D. R. (2002). Report of the Walkerton Inquiry: Part One: A Summary. In *Applied and Environmental Microbiology*.

Ojha, C. S. P., Hari Prasad, K. S., Ratha, D. N., & Surampalli, R. Y. (2012). Virus Transport through Unsaturated Zone: Analysis and Parameter Identification. *Journal of Hazardous, Toxic, and Radioactive Waste*, *16*(2), 96–105. https://doi.org/10.1061/(asce)hz.2153-5515.0000102

Ojha, C. S. P., Surampalli, R. Y., Sharma, P. K., & Joshi, N. (2011). Breakthrough Curves and Simulation of Virus Transport through Fractured Porous Media. *Journal of Environmental Engineering*, *137*(8), 731–739. https://doi.org/10.1061/(asce)ee.1943-7870.0000374

Paniconi, C., Aldama, A. A., & Wood, E. F. (1991). Numerical evaluation of iterative and noniterative methods for the solution of the nonlinear Richards equation. *Water Resources Research*, *27*(6). https://doi.org/10.1029/91WR00334

Pinon A, Vialette M (2019). Survival of viruses in water. Intervirology. 61(5):214–22.

Rajeev Gandhi, B. G., Bhattacharjya, R. K., & Satish, M. G. (2017). Simulation–Optimization-Based Virus Source Identification Model for 3D Unconfined Aquifer Considering Source Locations and Number as Variable. *Journal of Hazardous, Toxic, and Radioactive Waste*, *21*(2), 04016019. https://doi.org/10.1061/(asce)hz.2153-5515.0000334

Rao, S. V. N. (2006). A computationally efficient technique for source identification problems in three-dimensional aquifer systems using neural networks and simulated annealing. *Environmental Forensics*, *7*(3). https://doi.org/10.1080/15275920600840560

Ratha, D. N., Hari Prasad, K. S., & Ojha, C. P. (2009). Analysis of virus transport in groundwater and identification of transport parameters. *Practice Periodical of Hazardous, Toxic, and Radioactive Waste Management*, *13*(2), 98–109. https://doi.org/10.1061/(ASCE)1090-025X(2009)13:2(98)

Rogers, L. L., & Dowla, F. U. (1994). Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling. *Water Resources Research*, *30*(2), 457–481. https://doi.org/10.1029/93WR01494

Şahin, A. U. (2018). A Particle Swarm Optimization Assessment for the Determination of Non-Darcian Flow Parameters in a Confined Aquifer. *Water Resources Management*, *32*(2). https://doi.org/10.1007/s11269-017-1837-9

Samuel, M. P., & Jha, M. K. (2003). Estimation of Aquifer Parameters from Pumping Test Data by Genetic Algorithm Optimization Technique. *Journal of Irrigation and Drainage Engineering*, *129*(5). https://doi.org/10.1061/(asce)0733-9437(2003)129:5(348)

Schijven, J. F., & Hassanizadeh, S. M. (2000). Removal of viruses by soil passage: Overview of modeling, processes, and parameters. *Critical Reviews in Environmental Science and Technology*, *30*(1), 49–127. https://doi.org/10.1080/10643380091184174

Schijven, J. F., Hassanizadeh, S. M., & De Bruin, R. H. A. M. (2002). Two-site kinetic modeling of bacteriophages transport through columns of saturated dune sand. *Journal of Contaminant Hydrology*, *57*(3–4), 259–279. https://doi.org/10.1016/S0169-7722(01)00215-7

Schijven, J. F., Hoogenboezem, W., Hassanizadeh, S. M., & Peters, J. H. (1999). Modeling removal of bacteriophages MS2 and PRD1 by dune recharge at Castricum, Netherlands. *Water Resources Research*, *35*(4), 1101–1111. https://doi.org/10.1029/1998WR900108

Schijven, J. F., & Šimůnek, J. (2002). Kinetic modeling of virus transport at the field scale. *Journal of Contaminant Hydrology*, *55*(1–2), 113–135. https://doi.org/10.1016/S0169-7722(01)00188-7

Seetha, N., Mohan Kumar, M. S., & Majid Hassanizadeh, S. (2014). Modeling the co-transport of viruses and colloids in unsaturated porous media. *Journal of Contaminant Hydrology*, *181*, 82–101. https://doi.org/10.1016/j.jconhyd.2015.01.002

Shahrbanoo Hazrati, Y., & Datta, B. (2017). Self-organizing map based surrogate models for contaminant source identification under parameter uncertainty. *International Journal of GEOMATE*, *13*(36). https://doi.org/10.21660/2017.36.2750

Sharma, Y. C., Mukherjee, A. K., Srivastava, J., Mahato, M., & Singh, T. N. (2008). Prediction of various parameters of a river for assessment of water quality by an intelligent technique. *Chemical Product and Process Modeling*, *3*(1). https://doi.org/10.2202/1934-2659.1181

Shlesinger, M. F., Zaslavsky, G. M., & Frisch, U. (1995). Lévy Flights and Related Topics in Physics. *Proceedings of the International Workshop Held at Nice*.

Šimůnek, J., & Van Genuchten, M. T. (1996). Estimating unsaturated soil hydraulic properties from tension disc infiltrometer data by numerical inversion. *Water Resources Research*, *32*(9), 2683–2696. https://doi.org/10.1029/96WR01525

Šimůnek, J., Van Genuchten, M. T., & Šejna, M. (2012). Hydrus: Model use, calibration, and validation. *Transactions of the ASABE*, *55*(4).

Sim, Y., & Chrysikopoulos, C. V. (1995). Analytical Models for One-Dimensional Virus Transport in Saturated Porous Media. *Water Resources Research*, *31*(5). https://doi.org/10.1029/95WR00199

Sim, Y., & Chrysikopoulos, C. V. (1998). Three-dimensional analytical models for virus transport in saturated porous media. *Transport in Porous Media*, *30*(1). https://doi.org/10.1023/A:1006596412177

Singh, A., Kushwaha, S., Alarfaj, M., & Singh, M. (2022). Comprehensive Overview of Backpropagation Algorithm for Digital Image Denoising. *Electronics (Switzerland)*, *11*(10). https://doi.org/10.3390/electronics11101590

Singh, R. M., & Datta, B. (2004). Groundwater pollution source identification and simultaneous parameter estimation using pattern matching by artificial neural network. *Environmental Forensics*, *5*(3). https://doi.org/10.1080/15275920490495873

Singh, R. M., & Datta, B. (2006). Identification of Groundwater Pollution Sources Using GA-based Linked Simulation Optimization Model. *Journal of Hydrologic Engineering*, *11*(2). https://doi.org/10.1061/(asce)1084-0699(2006)11:2(101)

Siregar, S. P., & Wanto, A. (2017). Analysis of Artificial Neural Network Accuracy Using Backpropagation Algorithm In Predicting Process (Forecasting). *IJISTECH (International Journal Of Information System & Technology)*, *1*(1). https://doi.org/10.30645/ijistech.v1i1.4

Sreekanth, J., & Datta, B. (2011). Comparative Evaluation of Genetic Programming and Neural Network as Potential Surrogate Models for Coastal Aquifer Management. *Water Resources Management*, *25*(13). https://doi.org/10.1007/s11269-011-9852-8

Straub, T. M., Pepper, I. L., & Gerba, C. P. (1993). Hazards from pathogenic microorganisms in land-disposed sewage sludge. In *Reviews of Environmental Contamination and Toxicology* (Vol. 132). https://doi.org/10.1007/978-1-4684-7065-9_3

Swathi, B., & Eldho, T. I. (2018). Aquifer parameter and zonation structure estimation using meshless local Petrov-Galerkin method and particle swarm optimization. *Journal of Hydroinformatics*, *20*(2). https://doi.org/10.2166/hydro.2017.060

Tamer Ayvaz, M., & Gurarslan, G. (2019). A hybrid optimization approach for parameter estimation of confined and leaky confined aquifers. *Water Science and Technology: Water Supply*, *19*(8). https://doi.org/10.2166/ws.2019.117

Thompson, S. S., & Yates, M. V. (1999). Bacteriophage inactivation at the air-water-solid interface in dynamic batch systems. *Applied and Environmental Microbiology*, *65*(3). https://doi.org/10.1128/aem.65.3.1186-1190.1999

Tim, U. S., & Mostaghimi, S. (1991). Model for Predicting Virus Movement Through Soils. *Groundwater*, *29*(2). https://doi.org/10.1111/j.1745-6584.1991.tb00517.x

Torkzaban, S., Hassanizadeh, S. M., Schijven, J. F., & Van Den Berg, H. H. J. L. (2006). Role of air-water interfaces on retention of viruses under unsaturated conditions. *Water Resources Research*, *42*(12), 1–11. https://doi.org/10.1029/2006WR004904

UC Berkeley Technical Completion Reports Title Virus Transport Through Percolating Beds Publication Date. (n.d.).

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

Van Genuchten, M. Th., & Nielsen, D. R. (1985). On describing and predicting the hydraulic properties. *Annales Geophysicae*, *3*(5).

Vauclin, M., Khanji, D., & Vachaud, G. (1979). Experimental and numerical study of a transient, two-dimensional unsaturated-saturated water table recharge problem. *Water Resources Research*, *15*(5). https://doi.org/10.1029/WR015i005p01089

Vilker, V. L., & Burge, W. D. (1980). Adsorption mass transfer model for virus transport in soils. *Water Research*, *14*(7). https://doi.org/10.1016/0043-1354(80)90256-0

Wagner, B. J. (1992). Simultaneous parameter estimation and contaminant source characterization for coupled groundwater flow and contaminant transport modelling. *Journal of Hydrology*, *135*(1–4). https://doi.org/10.1016/0022-1694(92)90092-A

Wakode HB, Baier K, Jha R, Azzam R (2018). Impact of urbanization on groundwater recharge and urban water balance for the city of Hyderabad, India. *Int Soil Water Conserv Res. 6(1):51–62.*

Wan, J., & Wilson, J. L. (1992). *Colloid Transport and the Gas—Water Interface in Porous Media*. https://doi.org/10.1021/bk-1992-0491.ch005

Wellings, F. M., Lewis, A. L., Mountain, C. W., & Pierce, L. V. (1975). Demonstration of Virus in Groundwater after Effluent Discharge onto Soil. *Applied Microbiology*, *29*(6). https://doi.org/10.1128/am.29.6.751-757.1975

Wellings, F. M., Sather, G. E., & Hammon, W. McD. (1971). Immunoelectrophoretic Studies of the California Encephalitis Virus Group. *The Journal of Immunology*, *107*(1). https://doi.org/10.4049/jimmunol.107.1.252

Wildenschild, D., Hopmans, J. W., & Simunek, J. (2001). Flow Rate Dependence of Soil Hydraulic Characteristics. *Soil Science Society of America Journal*, *65*(1), 35–48. https://doi.org/10.2136/sssaj2001.65135x

Woodbury, A. D., & Rubin, Y. (2000). A full-Bayesian approach to parameter inference from tracer travel time moments and investigation of scale effects at the Cape Cod experimental site. *Water Resources Research*, *36*(1). https://doi.org/10.1029/1999WR900273

Woodbury, A. D., & Ulrych, T. J. (2000). A full-Bayesian approach to the groundwater inverse problem for steady state flow. *Water Resources Research*, *36*(8). https://doi.org/10.1029/2000WR900086

Xiao, C. N., Lu, W. X., An, Y. K., Gu, W. L., & Zhao, Y. (2015). Simulation-optimization model of identification of groundwater pollution sources based on two coupling method. *Zhongguo Huanjing Kexue/China Environmental Science*, *35*(8).

Xing, Z., Qu, R., Zhao, Y., Fu, Q., Ji, Y., & Lu, W. (2019). Identifying the release history of a groundwater contaminant source based on an ensemble surrogate model. *Journal of Hydrology*, *572*. https://doi.org/10.1016/j.jhydrol.2019.03.020

Yang, X. S. (2012). Flower pollination algorithm for global optimization. *In International conference on unconventional computing and natural computation* (pp. 240-249). Berlin, Heidelberg: Springer Berlin Heidelberg.

Yates, M. V. (1985). Septic Tank Density and Ground-Water Contamination. *Groundwater*, *23*(5). https://doi.org/10.1111/j.1745-6584.1985.tb01506.x

Yates, M. V., & Ouyang, Y. (1992). VIRTUS, a model of virus transport in unsaturated soils. *Applied and Environmental Microbiology*, *58*(5). https://doi.org/10.1128/aem.58.5.1609-1616.1992

Yates, M. V., & Yates, S. R. (1988). Modeling microbial fate in the subsurface environment. *Critical Reviews in Environmental Control*, *17*(4). https://doi.org/10.1080/10643388809388339

Yeh, H. Der, Lin, C. C., & Yang, B. J. (2014). Applying hybrid heuristic approach to identify contaminant source information in transient groundwater flow systems. *Mathematical Problems in Engineering*, *2014*. https://doi.org/10.1155/2014/369369

Yeh, W. W. (1986). Review of Parameter Identification Procedures in Groundwater Hydrology: The Inverse Problem. *Water Resources Research*, *22*(2). https://doi.org/10.1029/WR022i002p00095

Zhang, J., Zhang, X., Niu, J., Hu, B. X., Soltanian, M. R., Qiu, H., & Yang, L. (2019). Prediction of groundwater level in seashore reclaimed land using wavelet and artificial neural network-based hybrid model. *Journal of Hydrology*, *577*. https://doi.org/10.1016/j.jhydrol.2019.123948

Zhang, Q., Hassanizadeh, S. M., Raoof, A., van Genuchten, M. Th., & Roels, S. M. (2012). Modeling Virus Transport and Remobilization during Transient Partially Saturated Flow. *Vadose Zone Journal*, *11*(2), vzj2011.0090. https://doi.org/10.2136/vzj2011.0090

Zhao, C., Wang, Y., Chen, X., & Li, B. (2005). Simulation of the effects of groundwater level on vegetation change by combining FEFLOW software. *Ecological Modelling*, *187*(2–3). https://doi.org/10.1016/j.ecolmodel.2004.10.019

Zhao, Y., Qu, R., Xing, Z., & Lu, W. (2020). Identifying groundwater contaminant sources based on a KELM surrogate model together with four heuristic optimization algorithms. *Advances in Water Resources*, *138*. https://doi.org/10.1016/j.advwatres.2020.103540

LIST OF PUBLICATIONS

Das, Mamata, Rajib Kumar Bhattacharjya, and Suresh A. Kartha. "ANN-SFLA based parameter estimation method for an unsaturated–saturated simulation model." *Modeling Earth Systems and Environment* (2023): 1-15.

Das, Mamata, Rajib Kumar Bhattacharjya, and Suresh A. Kartha. " Estimation of Virus Transport Parameters in Unconfined Groundwater Aquifer: A Numerical Simulation and Flower Pollination Algorithm Approach." *Groundwater for sustainable development @Elsevier (Under review).*