

Hydrogeochemical controls and spatial variability of groundwater contamination in private wells
of the Great Bend Prairie aquifer, South-central Kansas

by

Thisura Chanuka Indralal Ilukgoda Gedara

B.S., University of Peradeniya, 2022

A THESIS

Submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Geology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2026

Approved by:

Major Professor
Matthew F. Kirk

Copyright

© Thisura Ilukgoda Gedara 2026.

Abstract

Private well water quality is largely unregulated in the United States, making it a responsibility of the well owner for testing and mitigating contamination. However, previous studies show that testing is infrequent, leaving well owners exposed to potential health threats. To help fill this knowledge gap, this study examined the quality of groundwater from private wells completed within the Great Bend Prairie Aquifer in south-central Kansas, a region characterized by extensive agricultural activity and sandy soils and subsurface sediment, which promote high subsurface permeability and increase vulnerability to contamination from the surface. Between 2020 and 2025, we sampled 217 wells in the study area which included a combination of domestic, irrigation, and livestock wells. Among potential threats to water quality, nitrate (NO_3^-) contamination was most common, with 46% of the wells having NO_3^- concentration above the US EPA maximum contaminant level for public water systems (10 mg/L as N) and concentrations ranging as high as 55.4 mg/L as N. Stable isotope data for a subset of samples indicated that chemical fertilizer was the main source of NO_3^- . Consistent with this result, our principal component analysis and random forest model indicated that NO_3^- concentrations were significantly controlled by hydrogeological, land use, geochemical, and climatic conditions, which are the main factors contributing to the transport of surface contaminants to the subsurface. Microbial NO_3^- reduction may also help limit NO_3^- accumulation, particularly along the northern margin of the study area where dissolved oxygen ($\text{O}_2(\text{aq})$) concentration was lower on average. Aside from NO_3^- contamination, trace element concentrations and concentrations of PFAS were generally low and below drinking water standards, but groundwater total dissolved solids (TDS) exceeded the US EPA secondary water quality standard (500 mg/L) in most samples and hardness was generally high. Nitrogen fertilizer

use may contribute to elevated TDS and hardness through acid production by nitrification in overlying crop soil and indeed, NO_3^- correlated with calcium (Ca^{2+}), magnesium (Mg^{2+}), hardness, and TDS. Additionally, a mixing analysis and water stable isotope data suggest that brine contamination from oil and gas production, which occurs over much of the aquifer, may have also increased the TDS of some of the groundwater we sampled. The proportion of wells with NO_3^- concentration above the standard for this study is high relative to comparable studies in other regions, even those focused on agricultural regions, demonstrating the importance of private well water quality testing, particularly in areas that are highly vulnerable to contamination from the surface. Moreover, study results demonstrate the value of private wells as opportunities to assess regional variation in water quality and controls on contaminant distributions.

Table of Contents

List of Figures	vii
List of Tables	ix
Acknowledgements	x
Chapter 1. Introduction	1
Chapter 2. Background Information	4
2.1 Potential contamination sources of private well water	4
2.2 Impacts of groundwater contamination on human health	7
2.3 Hydrogeology of the Great Bend Prairie Aquifer	8
Chapter 3. Materials and methods	11
3.1 Study area	11
3.2 Well selection	12
3.3 Field methods	13
3.4 Laboratory methods	16
3.5 Geospatial Analysis	17
3.6 Principal component analysis	21
3.7 Random forest model	23
Chapter 4. Results	27
4.1 Groundwater geochemistry	27
4.2 Stable isotope composition within private well water	35
4.3 NO ₃ ⁻ distribution and relationships to other water quality parameters	37
4.4 Relationship between NO ₃ ⁻ and hydrogeochemical parameters	39
4.5 Relationship of NO ₃ ⁻ concentration to land use and land cover conditions	41
4.6 Relationship of NO ₃ ⁻ concentration to soil and hydrogeological properties	43
4.7 Principal Component Analysis	44
4.8 Random forest analysis	46
Chapter 5. Discussion	48
5.1 Controls on NO ₃ ⁻ concentrations in the study area	48
5.1.1 Hydrogeochemical parameters	48
5.1.2 Microbial NO ₃ ⁻ reduction	50

5.1.3	Land use and cover (LULC)	51
5.1.4	Hydrogeological and environmental conditions	52
5.2	Principal component analysis (PCA)	54
5.3	Random forest model	56
5.4	Comparison of NO ₃ ⁻ contamination in private wells with other regions of the United States	58
5.5	Concluding remarks	60
Chapter 6. References		61
Appendix A – Random Forest Model		83
Appendix B – Principal Component Analysis		137

List of Figures

Figure 2.1: A) Bedrock geology underlying beneath the Great Bend Prairie aquifer and B) Vertical section from west to east across the region, showing the relation of the alluvial Great Bend Prairie aquifer to the underlying Cretaceous and Permian formations (Whittemore, 1993).....	9
Figure 3.1: Sampling setup based on the private well study by Holm et al. (2004). The 5-gallon bucket was used for flow measurement whereas the graduated cylinder was used for field parameter monitoring. For some wells, field parameters were measured with a dissolved oxygen meter in addition to a pH and temperature meter (shown). The sampling tube had a 0.375 inch internal diameter and consisted of low-density polyethylene (LDPE).	13
Figure 4.1: Box plot representing the variability of the anion data within the study area.....	29
Figure 4.2: Box plot representing the variability of the cation data within the study area	31
Figure 4.3: Piper diagrams showing grouped hydrogeochemical facies of groundwater in the study area: (a) Barton–Stafford–Rice (b) Reno–Kingman (c) Pratt–Kiowa–Edwards (Ca–HCO ₃ dominant), and (d) Pawnee.....	33
Figure 4.4: Distribution of trace element concentrations in groundwater samples from the study area.	34
Figure 4.5: The stable isotope composition plot $\delta^2\text{H}$ and $\delta^{18}\text{O}$ in private well water (The Global Meteoric Water Line (GMWL; Craig, 1961) and Local Meteoric Water Line (LMWL; Harvey & Welker, 2000) are shown for reference.)	36
Figure 4.6: isotopic composition values of $\delta^{15}\text{N}\text{--NO}_3^-$ and $\delta^{18}\text{O}\text{--NO}_3^-$ in private well water samples with reference fields from (C.-Q. Liu et al., 2006) for potential nitrate sources, including atmospheric deposition, NO_3^- fertilizers, NH_4^+ fertilizers, and manure	37
Figure 4.7: Map of NO_3^- contamination levels within the sampled wells.....	38
Figure 4.8: Relationships between NO_3^- concentration and selected geochemical parameters in groundwater samples.	40
Figure 4.9: Relationships between NO_3^- concentration (mg/L) and land use/land cover variables (m^2) within the study area.	42
Figure 4.10: Relationships between NO_3^- concentration (mg/L) and hydrogeological and climatic parameters	43

Figure 4.11: Scree plot showing variance explained by principal components, with PC1–PC4 accounting for ~75% of total variance.	45
Figure 4.12: PCA loading plot showing nitrate aligned with corn cultivation along PC1 and hydrogeological controls (Ksat, depth to groundwater, average precipitation) along PC2..	46
Figure 4.13: Variable importance in the random forest model	47
Figure 5.1: Cl ⁻ concentration versus Cl/Br mass ratio for groundwater and oil brine samples (Whittemore, 1993)	49
Figure 5.2: Percentage of sand within the soil (Data Source: UC Davies SoilWeb)	53
Figure 5.3: Nitrate contamination risk map developed using spatial prediction based on the random forest model	57

List of Tables

Table 3.1: Sample Types, Target Analytes, and Preservation Protocols for Water Quality Analysis.....	15
Table 4.1: Summarized statistics of the anion data within the study area	28
Table 4.2: Summarized statistics of the cation data within the study area	30
Table 4.3: Proportions of wells with NO_3^- concentrations below 3 mg/L as N, from 3-10 mg/L as N, and greater than 10 mg/L as N.	39

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor, Matthew Kirk, for his continuous guidance, support, and encouragement throughout this research project. His insight, constructive feedback, and motivation were very important. Besides my advisor, I want to acknowledge my committee members, Dr. Helene Avocat and Dr. Ganga Hettiarchchi, for their time and valuable contributions that strengthen this work. Further, I would like to thank all the professors, staff, and students in the Department of Geology at Kansas State University for their support and for creating a great academic atmosphere within the department.

Further, I must acknowledge financial support from the National Science Foundation (NSF), the Kansas Department of Health and Environment (KDHE), the Geological Society of America (GSA), and the Kansas Water Institute (KWI) at different stages of this project for its successful completion. Also, I want to extend my thanks to all the external laboratories and their staff members who helped us in analyzing our samples.

A special thank you to my wife, parents, and all my family members for their unconditional love and continuous support throughout my academic journey. Finally, I would like to thank all the wonderful people I met during my time in Manhattan.

Chapter 1. Introduction

Private wells can be identified as a primary source of drinking water for approximately 43 million people within the United States, mostly for residents in rural areas where public water systems are unavailable (DeSimone, 2009; Dieter et al., 2018; Lee & Murphy, 2020). It is estimated that 15.8 million private well systems exist, serving water for every one in five households across the country (Wallender et al., 2014). Regardless of the significance of private wells as a main source of drinking water, no proper federal-level regulation exists for maintaining the private well water quality, unlike for public water systems, which are regulated in the US by the Safe Drinking Water Act. Even though that authority has been devolved to the state level, private well water quality is not consistently regulated by states as well. (Bowen et al., 2019; Schmitt et al., 2024).

Private well owners are generally responsible for testing and maintaining their wells (Lee & Murphy, 2020), but in states that lack regulations, testing is limited. However, private well owners show little interest in regular monitoring of their well water due to a variety of reasons, such as ignorance, cost, privacy concerns, inconvenience of the testing processes, time issues, and lack of information associated with water testing locations, not having any health issues, not having any noticeable changes in aesthetic quality, and lack of knowledge about contaminations (Jones et al., 2006; Kreutzwiser et al., 2011). For example, (Malecki et al., 2017) conducted a statewide survey of 719 households in Wisconsin, USA, and found that one half of respondents reported testing within the last decade and only 10% within the past 12 months, despite state agency recommendations for annual testing. As a second example, (Swistock et al., 2013) surveyed 701 well owners in Pennsylvania and found that testing was rare, with 30% of respondents indicating they had never tested their water and 44% indicating it had only been

tested once. Therefore, the lack of proper monitoring and maintenance of private wells has the ability to expose well water users to a number of health issues linked to contaminated drinking water.

Groundwater extracted from these private well is vulnerable to contamination from both natural geochemical processes and anthropogenic activities, such as agriculture, septic systems, urbanization, and industrial inputs (P. Li et al., 2021; Rajurkar, 2025), making it highly susceptible to contaminants such as toxic metals, hydrocarbons, trace organic contaminants, pesticides, nanoparticles, microplastics, and other emerging contaminants, bacteria, viruses, and parasites (Craun et al., 2010; P. Li et al., 2021). Further, factors like improper private well designs, close proximity to contamination sources, poor well maintenance, and older well age can trigger the contamination associated with private wells (Stillo & MacDonald Gibson, 2017; Swistock et al., 2013). Even though groundwater experiences natural purification effects during recharge (Balke & Zhu, 2008), these processes are happening at a very slow rate, which takes decades or hundreds of years to vanish them entirely (P. Li et al., 2021). Therefore, proper regulation is always important for monitoring and maintenance of these private wells against these contaminants.

Due to the inconsistencies associated with private well water regulations, limited understanding of contamination occurrence, controlling factors, spatial variability, and associated human health risks. Even though numerous research studies are happening over many disciplines on private well water quality, their outputs are mostly confined to local or regional scales, as different geographical and geological areas face region-specific groundwater quality issues (Lee & Murphy, 2020). When coming to the state of Kansas, 71000 private wells exist within the state, providing water for household consumption and other domestic purposes (Ablah et al.,

2019) despite being unregulated by any of the state laws. Associated with this Great Bend Prairie Aquifer, several past studies have reported elevated NO_3^- levels and water quality degradation due to saline water mixing (Gardner & Sampson, 2024; Lane et al., 2020a). Most of these studies were based on either public water systems or monitoring wells, resulting a significant research gap over the water quality of private wells in south-central Kansas, exposing private well users to a significant health risk.

To address this significant research gap, this study evaluates the groundwater quality issues associated with private wells in south-central Kansas. In this study, we aimed to characterize the occurrence and spatial variability of the key contaminant within the study area while understanding the sources and control factors behind those contaminations through an integrated geochemical and geospatial approach.

Chapter 2. Background Information

2.1 Potential contamination sources of private well water

Intense agricultural activities can be identified as one of the main threats to the groundwater quality of private wells due to the extensive use of fertilizers and pesticides on croplands (Juan Carlos González García, 2025). They are causing direct as well as indirect impact on the rates and composition of groundwater recharge and biogeochemistry associated with aquifers underneath the cultivated areas. Direct effects include the movement of the excess inputs of fertilizers and herbicides into the aquifer. With this external supply of chemical constituents into the aquifer, it indirectly causes changes in the water-rock interactions within the soil (Böhlke, 2002). Through this direct and indirect contamination pathways, different types of inorganic contaminants such as nitrate (NO_3^-), chloride (Cl^-), sulfate (SO_4^{2-}), phosphate (PO_4^{3-}), potassium (K^+), calcium (Ca^{2+}), magnesium (Mg^{2+}), ammonium (NH_4^+), arsenic (As), uranium (U), selenium (Se) as well as pesticides, herbicides and different organic compounds can be contaminated to the groundwater (Böhlke, 2002; Hamilton & Helsel, 1995). Further, contamination can be caused by agricultural practices such as the treatment of farmlands with biosolids and animal manure. Biosolids have the potential to contaminate the soil and groundwater with emerging contaminants like per- and polyfluoroalkyl substances (PFAS) due to the absence of any federal regulation towards the PFAS concentrations in biosolids applied on farmlands (G. R. Johnson, 2022; Oviedo-Vargas et al., 2025; Wagner et al., 2015). The application of animal manure can also add nutrients and microbial pathogens to the groundwater (Kyte et al., 2023; McAllister & Topp, 2012).

Apart from crop cultivation, livestock operations also show a considerable contribution towards contaminations associated with private wells. According to the estimations, animal

feeding operations within the United States generate over 1.2 billion tons of waste annually (Palandri, 2025). However, the effects of the point source discharge associated with animal feeding operations on groundwater are spatially limited, extending only over an area of one to a few hectares (Harter et al., 2014). The main groundwater pollutants contained within the livestock wastes are N, P, K, Cl, and microorganisms (Goldberg, 1989; Kim & Kim, 2012). Apart from that, salts, trace elements, as well as lesser amounts of antibiotics, pesticides, and hormones can also be added to the environment (Copeland et al., 2006). Among these nitrogenous compounds are the dominant pollutants within animal waste (Goldberg, 1989). Nitrogen is present in livestock waste in both organic and inorganic forms, and, due to microbial action, it is converted to ammonia (NH_3), which is later turned into NO_3^- and NO_2^- . NO_3^- is the most soluble form of N within decomposed waste, and it readily migrates into the subsurface due to its soluble nature (Goldberg, 1989; Sahoo et al., 2016), polluting the aquifers below.

Further, pollution associated with on-site septic systems has emerged as a potential threat to the groundwater quality in the United States. These septic systems are soil-based wastewater treatment methods where the surrounding absorption area mainly carries out the physical, chemical, and biological wastewater treatment processes under aerobic and unsaturated conditions (Cogger, 1988). However, their impact varies from one system to another, depending on their localized contamination pathways (Mbae et al., 2024). They can act as a contamination source for a wide range of chemical contaminants and pathogens found within the groundwater. Coliform bacteria and NO_3^- are the most conventional types of contaminants associated with the majority of the septic systems (Schneider et al., 2016). In addition to that, chemical contaminants like phosphate, organic wastewater compounds like pharmaceuticals, personal care products, caffeine, and organophosphate flame retardants, PFAS compounds like perfluorosurfactants,

human hormones like drostenedione, estrone, and progesterone (Cogger, 1988; Conn et al., 2009; Schaider et al., 2014; Swartz et al., 2006) can be found associated with these on-site septic systems.

Industrial activities like oil and gas production also contributed to contamination in private wells through different pathways, such as surface spills and leaks during storage and transport, infiltration from unlined disposal pits, and migration along improperly constructed or degraded well casings (Hammond et al., 2020; Jackson et al., 2014). The disposal of both conventional and unconventional produced water containing salts, heavy metals, volatile organic compounds, and naturally occurring radioactive materials acts as the primary source of contaminants (Cantlay et al., 2025; Gannon et al., 2025) from oil production. Compared with produced water from conventional oil and gas wells, produced water from unconventional wells can be contaminated with potentially hazardous chemicals such as arsenic, selenium, strontium, various alcohols, chlorinated species, and BTEX (benzene, toluene, ethylbenzene, and xylene) compounds derived from the fracturing fluids (Fontenot et al., 2013; Hildenbrand et al., 2016).

The mining industry also causes contamination of groundwater extracted from private wells through the generation of acid mine drainage (AMD) (Madhav & Singh, 2021). Acid mine drainage is simply acidic and inorganic contaminant-rich wastewater that forms associated with active or abandoned mines through the oxidation of sulfide-bearing minerals under the influence of atmospheric oxygen and water (Anekwe & Isa, 2023; Nguegang & Ambushe, 2024), and it may concentrated with (Al), iron (Fe) and manganese (Mn), zinc (Zn), copper (Cu), nickel (Ni) and arsenic (As), chromium(Cr), radioactive elements such as uranium (U) and ions like Arsenate (AsO_4^{3-}) (Baena-Moreno et al., 2020; Nguegang & Ambushe, 2024). Apart from that, byproducts of mining operations, such as tailings, dust, waste rock, and wastewater, can cause

infiltration and diffusion of heavy metals, trace elements, nonmetals, and radionuclides, polluting groundwater (Chellam & Clifford, 2002; Hao et al., 2024). Under the influence of these kinds of processes, groundwater extracted from the mining-impacted regions is showing contamination with pollutants that are harmful to human health.

2.2 Impacts of groundwater contamination on human health

Nitrogen compounds can be identified as the most widespread contaminant within the subsurface, while being the reason behind several human diseases (Liu et al., 2011). Blue baby syndrome, or Methemoglobinemia, is caused when the NO_3^- concentration exceeds 10 mg/L as N, and the EPA's current maximum contaminant level (MCL) has been set to protect babies under the age of 3 months from it (Adelana, 2005). But the chronic exposure to NO_3^- can result in many other health issues like hypertension, increased infant mortality, central nervous system birth defects, diabetes, spontaneous abortions, respiratory tract infections, Balkan nephropathy, slowing of motor reflexes in children, NO_3^- esters dependence, and changes to the immune system, as well as bladder, colorectal, and gastric cancers (Adelana, 2005; Bruning-Fann & Kaneene, 1993; Fewtrell, 2004; Wadgaonkar, 2024).

Other inorganic contaminants are also causing some health effects, even though they are not as severe as NO_3^- . Excess F^- contamination can cause fluorosis, adversely affecting the liver, kidneys, muscles, heart, lungs, blood, and hormonal functions (Arif et al., 2012). SO_4^{2-} is considered to be non-toxic in low concentrations, but exposure to extremely high sulfate levels can cause gastrointestinal disturbances such as diarrhea, dehydration, and discomfort in some individuals (Preethi et al., 2025). Excess Sodium(Na^+) ion intake can contribute to heart failure to a certain extent, but the effect mostly depends on the degree of sodium retention within the

body (Proger et al., 1942). Calcium(Ca^{2+}) and magnesium(Mg^{2+}) contribute to the water hardness, but there are no proven health effects from the consumption of contaminated drinking water, even though there are some discussions over their contribution towards cardiovascular diseases (World Health Organization, 2009).

Metal-contaminated groundwater can cause different types of health issues in humans. Overall, the accumulation of heavy metals within the human body can cause severe health damage to the vital organs, such as the brain, liver, bones, cardiovascular organs, and kidneys as well as growth issues, organ development issues and negative pregnancy outcomes (Meena et al., 2020; Sankhla, 2019; Sinha & Prasad, 2020; Smith & Steinmaus, 2011). Further, heavy metal accumulation in water is identified as a serious threat to human health (Wadgaonkar, 2024), and these different metals cause different cancers. As an example, arsenic(As) accumulation causes cancers associated with skin, lung, bladder, liver, and urogenital organs, and lead(Pb) can result in cancer types like lung, kidney, stomach, and brain(Sinha & Prasad, 2020; Wadgaonkar, 2024).

Emerging contaminants like PFAS can also cause health issues like increased risk of metabolic syndromes such as type 2 diabetes, obesity, metabolic syndrome, thyroid diseases, disrupt hormonal balance, suppressed immune response, lowered infant birth weight, and impacts on liver function, as well as various cancers (Tsang et al., 2026; Xu et al., 2023).

2.3 Hydrogeology of the Great Bend Prairie Aquifer

This Great Bend Prairie aquifer is mainly composed of unconsolidated deposits of the Pliocene and Pleistocene ages, extending up to a maximum depth of 360ft (Fader & Stullken, 1978). These layers are mainly made up of clay, silt, and gravel with some layers of loosely cemented sand and caliche, which are assumed to be deposited by wind and the ancestral

Arkansas River. These sediments are capped by loamy and sandy mollisols and alfisols (Dodge et al., 1978) and floored by Permian-age rock in the east and Cretaceous-age rocks on the western side. The main bedrock types are sandstone, siltstone, and shale (Fig. 2.1). Within some

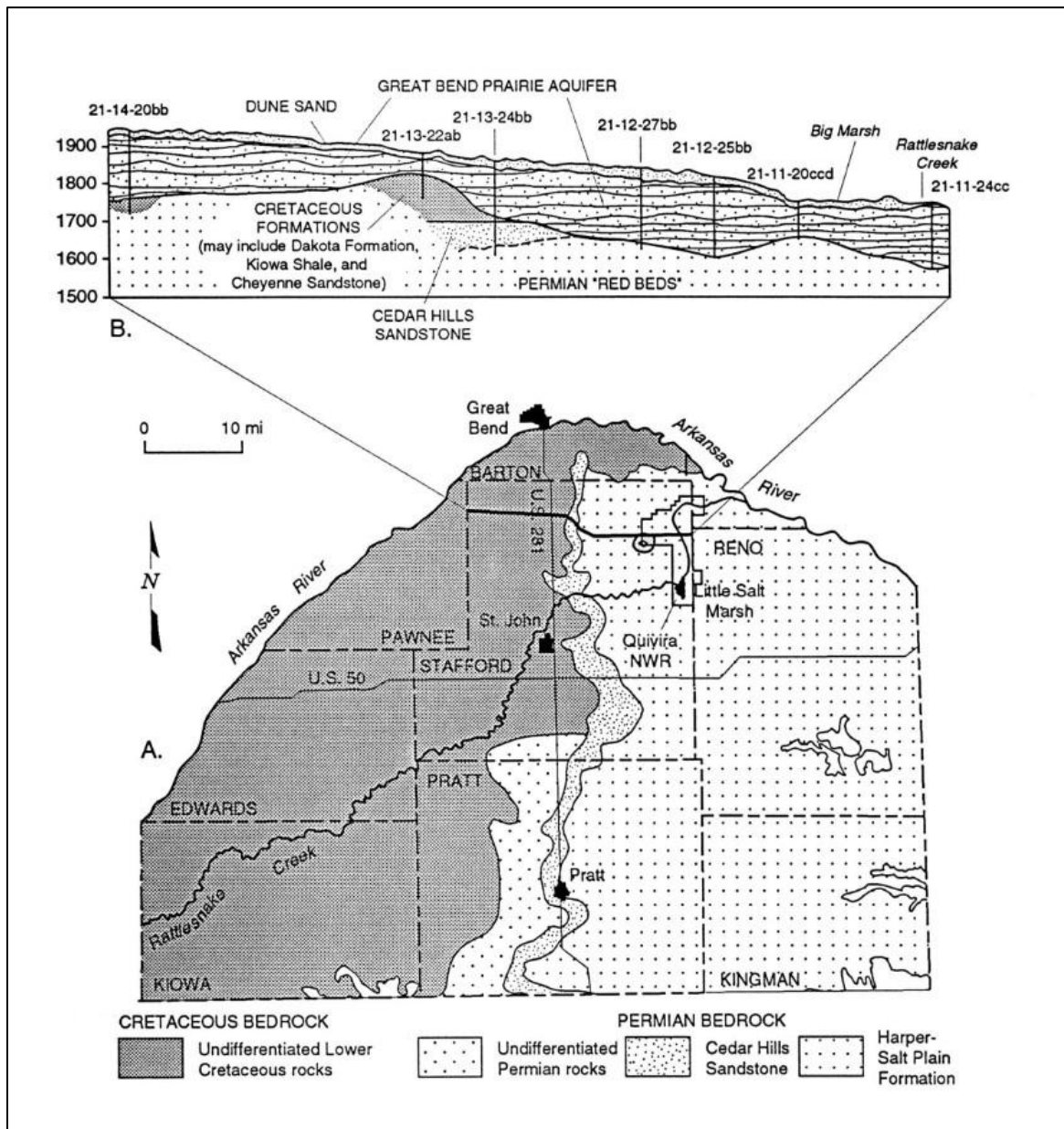


Figure 2.1: A) Bedrock geology underlying beneath the Great Bend Prairie aquifer and B) Vertical section from west to east across the region, showing the relation of the alluvial Great Bend Prairie aquifer to the underlying Cretaceous and Permian formations (Whittemore, 1993)

parts of the Permian bedrock, salt beds or salt-cemented layers can be found (Fader & Stullken, 1978; Macfarlane & Ackerman, 1983).

Apart from the undifferentiated Pleistocene deposits, which can be identified as the unconsolidated aquifer, Permian-age red beds also serve as aquifer units within the subsurface, and they are composed of reddish-brown siltstone, shale, and sandstone with lesser amounts of salt, gypsum, anhydrite, limestone, and dolomite. This bedrock unit can extend to a maximum thickness of 1270 feet. The Cretaceous rock layers overlie the Permian red beds and mainly consist of interbedded sandstone and shales, serving as the upper confining layer for red beds due to their aquitard nature (Cobb, 1981; Layton & Berry, 1973).

Chapter 3. Materials and methods

3.1 Study area

The Great Bend Prairie Aquifer is geographically positioned in the south-central portion of Kansas, spanning over Kiowa, Kingman, Pratt, and Stafford counties, along with considerable portions of Barber, Barton, Edwards, Pawnee, Reno, and Rice counties (Fader & Stullken, 1978). The South-central Kansas region is one of the main agricultural regions within central Kansas, with dominant crop types like corn, wheat, soybean, and sorghum (Sophocleous, 2010; Suttles et al., 2024). Compared to the rest of the region within Kansas, the soil is more sandy in nature and considered to be the greatest sand sheet in the area (Arbogast, 1996). This is evidence of a higher permeability, facilitating an effective transport of surface contaminants into the subsurface (Shapiro & Falcone, 2022).

The Great Bend Prairie aquifer is recharged mainly through direct infiltration of precipitation on the surface and underflow that happens laterally from west to east. In addition, upward leakages are possible from the bedrock as well (Fader & Stullken, 1978). Compared with the rest of the regions of the High Plains aquifer within Kansas, the Great Bend Prairie Aquifer is showing small declines and increases in water level due to relatively higher levels of precipitation and shallow depth to the water (Lewis Jr, 2025; Whittemore, 1993).

Water quality is the main issue associated with this aquifer, over water quantity. (Lane et al., 2020) study, which was carried out targeting 22 monitoring wells located within the Great Bend Prairie aquifer, has found that 7 of the wells are exceeding the MCL for NO_3^- s. This is mostly due to the application of fertilizer on farmlands and the disposal of animal waste. Several past studies have identified NO_3^- as a widespread contaminant within the Great Bend Prairie aquifer (Lane et al., 2020a; Townsend & Young, 1995a). Apart from that, groundwater within

the Great Bend Prairie aquifer can be contaminated through processes such as oil and gas production, landfills, municipal wastewater lagoons, and industrial wastes (Lane et al., 2020a; Lewis Jr, 2025). Apart from that, this aquifer is susceptible to natural saline water contamination due to the leaching of brine from the underlying Permian bedrocks (Sophocleous, 2010).

3.2 Well selection

During the period between 2020 and 2025, we sampled 217 private wells distributed across much of the Great Bend Prairie Aquifer. Most of the wells were domestic wells, although we also sampled a few irrigation wells, livestock wells, and wells used for mixed purposes.

Over the course of the study, we carried out multiple rounds of sampling, with each round focused on a particular subregion of the study area, and we used multiple approaches to connect with private well owners. Some wells were selected by visiting well owner residences and asking if they were willing to allow us to sample, but for most wells, we contacted well owners in advance to arrange sampling. In the latter case, to identify potential wells, we relied on the Kansas Geological Survey (KGS) WWC5 well record database as well as rural directories.

We attempted to sample wells distributed evenly across the study area. We did indeed cover much of the aquifer. However, sampling density is not uniform. In part, this reflects the fact that wells are not evenly distributed across the study region, but also that connecting with well owners was challenging. Most well owners we did successfully reach were willing to allow sampling in exchange for an assessment of their water quality. Furthermore, some well owners had multiple wells available to sample or helped put us in contact with other well owners for sampling.

3.3 Field methods

This study aimed to understand variation in water quality throughout the Great Bend Prairie Aquifer. To that end, we sampled wells from outside hydrants and spigots, avoiding treatment practices such as water softening and point of use reverse osmosis that are generally employed only for indoor water uses. Moreover, we used standard well purging methodology (ASTM, 2012) prior to collecting any water samples. These steps helped ensure that the water chemistry variation we observed reflects variation in the aquifer with limited variation added from the conditions within each water system.

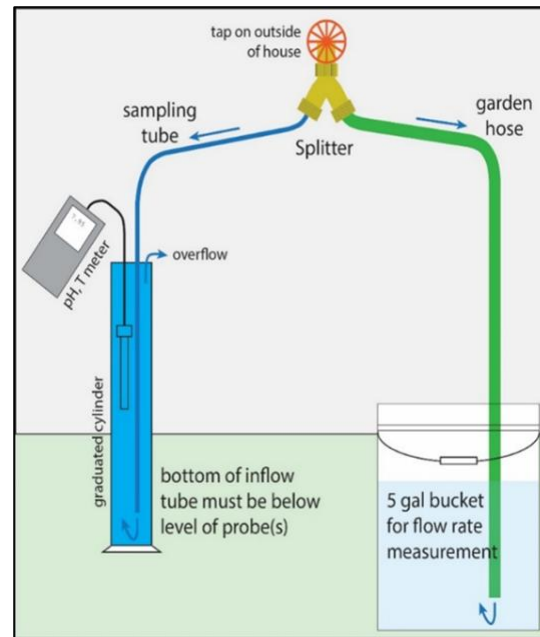


Figure 3.1: Sampling setup based on the private well study by Holm et al. (2004). The 5-gallon bucket was used for flow measurement whereas the graduated cylinder was used for field parameter monitoring. For some wells, field parameters were measured with a dissolved oxygen meter in addition to a pH and temperature meter (shown). The sampling tube had a 0.375 inch internal diameter and consisted of low-density polyethylene (LDPE).

For all wells sampled from 2023 through 2025 (156 wells overall), we initiated sampling by attaching a brass two-way splitter to the sampling hydrant or spigot, with one side connected

to a garden hose and the other to a small polyethylene tube with an inside diameter of 0.375 inches (Fig. 2.1). We then turned on the flow and measured the flow rate by recording time necessary to fill a 5-gallon bucket. Without adjusting the overall flow rate, we then opened the splitter valve to allow some flow through the sampling tube to begin monitoring field parameters, including pH and temperature for most of the wells and dissolved oxygen ($O_2(aq)$) and conductivity for a portion of the wells.

We monitored pH and temperature with Oakton pH 150 meters equipped with an Oakton double-junction, gel-filled electrode with automatic temperature compensation (ATC). We used YSI Pro2030 meter with a galvanic sensor to measure $O_2(aq)$ concentration, temperature, conductivity, and specific conductivity. We were not able to measure $O_2(aq)$ and conductivity for all wells because we did not have sufficient meters for all of the student teams who were collecting samples during each field trip. Before collecting any measurements each day, we calibrated all the field meters using three-point calibration for pH and single-point calibrations for $O_2(aq)$ and conductivity.

During the field measurements, we placed field meter probes within a 1 L polypropylene graduated cylinder (Fig. 2.1), ensuring that they were positioned above the level of the inflow tube. As such, the graduated cylinder functioned as a flow-through cell, which limited interaction of the groundwater with the atmosphere prior to the measurements. We recorded the first reading after allowing flow through the sampling tube and cylinder for five minutes. Then we recorded two more reading sets with five minutes interval between each.

After completing field measurements, we reduced the flow through the garden hose, maintaining it mainly through the small polyethylene tube, and recorded the time to identify how long the well was purged. The pumping rate together with the purge time could then be used to

evaluate the purge volume, which was on average 195 gallons (733.4 L) for wells with complete data recorded. Next, we attached a 0.45 μm syringe filter to the tube and began filling ample bottles except those for PFAS and NO_3^- isotopes. We used 33 mm diameter nylon membrane disk filter (Fisherbrand) for most wells, but if the filters rapidly clogged, we used high-capacity Sterivex filters (MilliporeSigma).

Sample storage and preservation are described in Table 3.1. All sample bottles were acid-washed and thoroughly rinsed with 18 M Ω deionized water prior to sampling, except the water isotope tubes, NO_3^- stable isotope bottles, and PFAS bottles. We also rinsed each bottle with sample prior to filling the bottle. If the sample being collected was intended to be filtered, the rinse was with filtered water.

Sample Description	Analytes	Bottle type	Filter (Y/N)	Preservation Process
Anions	alkalinity, F^- , Cl^- , NO_2^- , Br^- , NO_3^- , SO_4^{2-}	60 mL HDPE	Yes	Cold
Cations	NH_4^+ , Na^+ , K^+ , Mg^{2+} , Ca^{2+} , Sr^{2+}	30 mL HDPE	Yes	Cold 0.2% HNO_3
Trace Elements	As, Se, U, Mn, and Fe	125 mL HDPE	Yes	Cold 0.2% HNO_3
NPOC/TN	dissolved organic C, total dissolved N	60 mL amber glass	Yes	Cold 0.2% HCl
Water stable isotopes	$\delta^2\text{H-H}_2\text{O}$, $\delta^{18}\text{O-H}_2\text{O}$	2 mL glass tube	Yes	Cold
NO_3^- stable isotopes	$\delta^{15}\text{N-NO}_3^-$, $\delta^{18}\text{O-NO}_3^-$	1 L HDPE	No	Cold in field /frozen in lab
PFAS	PFAS compounds	250 mL (Qty: 2)	No	Cold

Table 3.1: Sample Types, Target Analytes, and Preservation Protocols for Water Quality Analysis

We stored samples on ice in coolers and transferred them to refrigerators or a freezer after returning to the lab. At the laboratory, we acidified cation and trace elements samples with

trace element grade nitric acid, and non-purgeable organic carbon (NPOC)/total nitrogen (TN) samples with reagent grade hydrochloric acid.

One exception to the above sampling procedure is for samples collected during 2020 and 2021. For those samples, we did not have the sampling apparatus shown in Figure 2.1, though we did use a comparable purging strategy. Additionally, for some wells we had to filter water from beakers using 60 mL syringes because the water was produced from discharge tubes that did not have threaded fittings.

During multiple field trips, we collect sets of field blanks as part of our quality control process. For field blanks, we filled sample bottles with 18 M Ω deionized water directly (PFAS) or using syringes fitted with the same filters used for groundwater samples (all other analytes).

3.4 Laboratory methods

After finishing fieldwork, we began analyzing the samples within less than 3 days to avoid any physicochemical alteration. All the analytical procedures were carried out following the EPA guidelines to make sure that the dataset is produced with high precision and accuracy.

Alkalinity was determined using an endpoint titration method either manually with a buret or with a Thermo Scientific Orion T910 automatic titrator. We used end point titrations with 0.02 N H₂SO₄.

For the anion and cation analysis, we utilized the Thermo Scientific Dionex ICS-1100 ion chromatograph (ICs) system following EPA 300.0 and 300.7, respectively. Nonpurgeable organic carbon (NPOC) and total nitrogen (TN) analysis was done using a Shimadzu TOC-L instrument. All these facilities were available in the Department of Geology at Kansas State University.

We analyzed trace element samples from Olathe public works using the PerkinElmer Nexion 350X ICP-MS instrument following EPA Method 200.8. In each sampling round, we analyzed those samples within a time frame of less than 6 months, which is the recommended time frame to store those samples.

The stable isotope ratios of water ($\delta^{18}\text{O}$ and δD) were measured using a Picarro L-i2130 water analyzer, available in the Department of Biology at Kansas State University, with 0.1 and 0.5 ‰ precision, respectively.

For the PFAS analysis, we send our samples to the Department of Chemistry, Kansas State University, and analyze them using the Waters Xevo TQS micro with a 2D ACQUITY UPLC (UPLC-MS/MS triple quad) instrument, following the US EPA method 1633 SPE.

3.5 Geospatial Analysis

For all map preparation and spatial analysis, we used ArcGIS Pro (Version 3.4) and its built-in geoprocessing tools. Here, we downloaded the necessary raster and shapefile data from various open spatial data sources on the internet.

In order to assess the influence of land management in our analysis, we used the land use and land cover (LULC) data and the Cropland Data Layer (CDL) published by the United States Geological Survey (USGS). The Great Bend Prairie Aquifer region is identified as a region that grows corn as the most common crop and generally does not follow crop rotation practices that switch to legumes or less fertilizer-demanding crops while using rye or wheat as common winter cover crops (Townsend et al., 1996). Recognizing this variability in annual LULCs, we assumed that agricultural practices within a given field remain relatively stable and simplified our analysis by considering only the LULCs for the years 1990, 2000, 2010, and 2015. However, it was not

possible to follow the same procedure for CDLs, as they have been published only since 2008. For CDL, we considered only data for the years 2010 and 2015.

For different LULCs and CDLs, different circular buffer sizes of 500m, 750m, 1000, 2000m were considered, assessing the contribution of the land use and land cover types in the surrounding area and longer groundwater travel times (Wheeler et al., 2015). It has been suggested that a 10-year travel time for groundwater is a reasonable one for a flat region with well-drained soil (Dubrovsky et al., 2010). In order to initiate the data extraction process from LULCs and CDLs, new layers with different buffer sizes around sampled wells were created using the “Buffer” tool, and then the " Zonal Statistics as Table" tool was run to extract the total amount of different land use and land cover or crop types within a considered buffer radius.

Apart from the LULC and crop types, total fertilizer loading and the annual average fertilizer usage were also considered. Here, two different types of approaches were used. Initially, the annual county-level fertilizer estimate data published for the periods from 1987 to 2012 were used to assess the total and average nitrogen (N) fertilizer loading since 1980 (Brakebill & Gronberg, 2017). The second approach was to account for crop-based N fertilizer usage (kg) based on the annual US averages published by the National Agricultural Statistics Service (NASS) agricultural chemical use program (*USDA - National Agricultural Statistics Service - Surveys - Agricultural Chemical Use Program*, 2026) to utilize more site-specific data with better resolution in the statistical analysis. For the crop types that lack any annual US-based fertilizer usage averages, recommended fertilizer guidelines (lbs/ac) were used.

Soil properties were incorporated into our spatial analysis as they are some key controlling factors influencing the infiltration, retention, and transport of NO_3^- and other contaminants through the subsurface (Burow et al., 2010; Hamlin et al., 2022). Most of the soil-

associated parameter data were downloaded from the "SoilWeb" database maintained by the California Soil Resource Lab of the University of California, Davis. We used the soil sand percentage, saturated hydraulic conductivity, drainage class, soil depth, soil order, pH, soil organic matter, and electrical conductivity data downloaded from their open web database. As the impacts of soil characteristics on groundwater movements and associated processes are not limited exactly to the well location, we calculated the mean value within a 1km buffer from the well location for parameters such as soil sand percentage, saturated hydraulic conductivity, soil organic matter, and electrical conductivity.

Using the High Plains aquifer properties dataset published by the Kansas Geological Survey (KGS), data such as national elevation data, bedrock type, bedrock depth, predevelopment water table elevation, predevelopment saturated thickness, water table elevation (2022-2024), amount of predevelopment water level changes (2022-2024), and percentage of predevelopment water level changes (2022-2024) were extracted for the well location.

Precipitation has a direct effect on NO_3^- contamination in groundwater (Sun et al., 2024), and in order to assess its relationship, precipitation data were extracted for the study area from the PRISM weather data website maintained by Oregon State University. Here, a set of 400 random points for the study area was created using the "create random points" tool, and the 30-year average precipitation data were extracted for those points from the PRISM database. Then the Empirical Bayesian Kriging (EBK) tool available in ArcGIS PRO software was used to create a continuous raster surface for the downloaded precipitation data values in the study area, and values for each specific location were extracted using that layer.

Distance to the streams was calculated using the "Kansas Streams" data file downloaded from the Kansas Geoportal, and it is a section of data extracted for the state of Kansas from the

National Hydrography Dataset (NHD) representing the order 1-9 streams. We used the “Euclidean Distance” tool to calculate the distance to the closest streams in meters (Atabati et al., 2022; *Euclidean Distance (Spatial Analyst)—ArcMap | Documentation*, 2021). Here, we calculated the distance to the closest streams of order 1-2 and order 3-9, as well as the distance to any of the streams of order 1-9 separately.

Considering the influence from human activity (Zhou et al., 2025), apart from land use and agriculture, the distance to the closest urban area from each specific well was calculated using the “Tiger 2020 Census Designated Places” dataset downloaded from the Kansas Geoportal. Here, the same approach that was employed to calculate the distance to the rivers was followed using the “Euclidean Distance” tool (*Euclidean Distance (Spatial Analyst)—ArcMap | Documentation*, 2021).

The certified animal feeding operations shapefile published by the Kansas Department of Health and Environment (KDHE) was used to assess the impact of those on NO_3^- contamination. Here, the animal type and size of the operations were standardized to simplify the analysis. For that, animal counts in each of the feeding operations were converted to general Kansas animal units following the classification of the Kansas Department of Health and Environment (KDHE, 2017). Instead of using the nearest distance tool, here we assess the effect of both size and distance from the animal feeding operations using the “Kernel density” tool, which performs a Gaussian kernel function (Zirkle et al., 2016).

The Kansas surface geology map, administrative boundaries, and most of the other basic geospatial data required for the spatial analysis were downloaded directly from the Kansas Geoportal Website. After processing the final spatial dataset, a numerical dataset was exported to CSV files to use in the statistical analysis.

3.6 Principal component analysis

Principal Component Analysis (PCA) is a statistical technique that is used to reduce the dimensionality of a large dataset while increasing interpretability and minimizing information loss (Jolliffe & Cadima, 2016). In a system with multiple input variables, this technique is useful in determining the variables with the greater eigenvalues to delineate the core of the problem (Chen et al., 2022). With its effectiveness in determining the natural and anthropogenic influences on water quality, this statistical technique is widely used in hydrogeochemical studies (Helena, 2000; Shrestha & Kazama, 2007).

PCA is performed using the correlation matrix, which is appropriate for standardized datasets (Jolliffe & Cadima, 2016). To derive the principal components (PCs), eigenvalues and eigenvectors are computed (Rao, 2014). Each of these PCs represents a linear combination of original variables. The number of significant components retained for the final interpretation was determined based on the Kaiser criterion (eigenvalues > 1) and cumulative variance explained (Cloutier et al., 2008).

Variable selection for the PCA analysis was done using several approaches. Here, we utilized the outcomes from the pairwise correlation analysis to assess which variables needed to be essentially included. However, most of those relationships are weak or sometimes not statistically significant, making it difficult to set a threshold for variable selection. Therefore, we also included some variables that did not strongly correlation but have been identified as key parameters contributing to NO_3^- contamination in previous studies. In order to ensure that the selected variables were appropriate to use in a PCA the Kaiser–Meyer–Olkin (KMO) and Bartlett's tests were carried out. If the KMO value is below 0.50 for any variables, those were considered unsuitable for PCA. Any dataset with an overall KMO value greater than 0.60

alongside a significant Bartlett's test ($p < 0.05$) can be considered to be appropriate to use in a multivariate dimension reduction technique like PCA (Kheirfam & Vafakhah, 2015; McLeod et al., 2017; “On Bartlett’s Sphericity Test for Correlation,” 2022; Vafakhah & Janizadeh, 2021).

The final combination of variables for the analysis included NO_3^- concentration, soil organic matter (SOM), cultivated land, herbaceous land cover, saturated thickness, saturated hydraulic conductivity (Ksat), average annual precipitation, corn land cover, water table elevation, and distance to streams.

To prepare data for the analysis, the skewness of the selected variables was addressed by carrying out appropriate transformations, as PCA is not a suitable technique to use with such data (Maadooliat et al., 2015). For the variables with highly right-skewed data, log transformation was introduced (feng et al., 2014), and for data with moderate right-skewness, square-root transformations were applied (Bakker, 2026; *R: Common Data Transformations and Standardizations*, 2025). Additionally, missing data was handled by using listwise deletion, as the number of missing observations containing one or more missing values was very low and those were entirely excluded from the PCA (Kang, 2013).

As these variables are originally measured on different scales, standardizing data was essential prior to the PCA (Jolliffe & Cadima, 2016), and it was carried out internally by the `prcomp()` function in R. The computational and visualization component of the PCA was done with the help of RStudio software and its “factoextra” and “tidyverse” packages. Eigenvalues, the proportion of explained variance, principal component loadings, and principal component scores were calculated for interpretation. Scree plots, biplots, correlation circles, and variable contribution plots were generated for better visualization of the outcome.

3.7 Random forest model

A random forest model is an ensemble machine learning approach that uses a combination of decision tree predictors (Breiman, 2001) to improve prediction performance by reducing the overfitting and instability associated with a single decision tree (He et al., 2022). A single decision tree is grown by recursively partitioning the dataset through binary splits until a previously specified stopping condition is reached. In a random forest model, multiple decision trees are constructed using bootstrap samples selected from the training dataset, while a random subset of predictor variables is selected at each split. Then the outcome of each decision tree is aggregated into a final output using voting or averaging for classification or regression problems, respectively (Rigatti, 2017; Speiser et al., 2015, 2019).

Here, we employed random forest classification to identify areas with higher and lower contamination risk, and the sampled wells were reclassified as either 0 (low risk) or 1 (high risk) (Sarkar et al., 2022) by setting the EPA standard for NO_3^- in groundwater (10 mg/L as N) as the classification criterion.

The random forest classifier was developed using the “ranger” package in R, which provides a computation-efficient implementation suitable for high-dimensional datasets like environmental data. This package has the capability to support ensembles of classification, regression, and survival trees (Wright & Ziegler, 2017). Our groundwater contamination model was developed using the probability estimation option implemented in the ranger package, allowing each prediction to represent the probability of classifying each sampled well into either low or high contamination groups (Kuhn & Johnson, 2013).

In order to evaluate model performance, the dataset was divided into training (70%) and testing (30%) subsets (Sivakumar et al., 2024) using a stratified training approach to preserve

similar proportions of high- and low-risk samples within both training and testing datasets. A total of 100 independent training datasets were generated using different random seeds, and the performance of the model was evaluated for each of the repetitions. This repeated validation process provides a more reliable estimate of the model performance than depending on a single train-test split (Schader et al., 2024). All the missing predictor values within each training dataset were treated with training data medians (Ranganathan & Hunsberger, 2024).

An automated tree convergence analysis was carried out by assessing the out-of-bag (OOB) predictions to determine the optimal number of trees (num.trees) required to obtain the optimal model performance (Probst & Boulesteix, 2018). The smallest number of trees producing an OOB AUC within 0.005 of the maximum observed value was selected.

The predictive performance of a random forest depends on several hyperparameters that control how the individual decision trees are generated. Here, we adjusted the hyperparameters, such as the number of predictor variables at each split (mtry), and the minimum number of observations in a terminal node (min_node_size) (Probst et al., 2019). With the intention of improving model development efficiency, an automated process was followed to tune the hyperparameters.

Following the repeated validation, a final reproducible random forest model was developed using a fixed random seed and the hyperparameters that were identified in the previous steps.

We evaluated the model performance using several metrics, including ROC AUC, overall accuracy, sensitivity, specificity, balanced accuracy, and precision (Cabot & Ross, 2023; Kuhn & Johnson, 2013). In order to quantify model stability and robustness, mean, standard deviation,

median, minimum, and maximum values of each performance metric were calculated across the 100 repeated training-test splits.

At the end of this repeated validation process, a final reproducible random forest model was developed with the help of a fixed random seed. Here, the final model was developed using the optimum tuning parameters identified from the previous steps. Then, the variable importance of the final model was quantified using permutation importance, which is calculated by measuring the decrease in prediction accuracy after randomly permuting each predictor variable.

For the development of the contamination risk map, all the predictor raster layers were standardized to a common coordinate system (NAD 1983 UTM Zone 14N) and spatial resolution (30 m × 30 m) before proceeding with the spatial predictions. The prediction map was developed only up to the extent of the dissolved oxygen raster, which had the smallest spatial coverage among all predictor rasters. That layer was set as the reference for the masking, and all the data beyond that spatial extent of the remaining layers were excluded. Then each of the raster layers was aligned to a common template using bilinear interpolation to confirm that all layers have identical cell size, coordinate system, origin, and spatial extent (Hijmans et al., 2020). After the alignment of all the raster layers, missing pixel values in any of the raster layers were treated with corresponding median values calculated from the final model training dataset. The aligned raster layers were finally stacked in the same variable order as they were used to train the random forest model (Cherlinka et al., 2025), which predicted the probability for each 30 x 30 m grid cell in the study area to exceed the NO_3^- concentration beyond the EPA standard for drinking water (10 mg/L as N).

Using this probability raster generated based on the random forest model, a binary contamination risk map was generated using the optimal probability threshold determined based

on the Youden Index. Grid cells with predicted probabilities equal to or greater than the optimal threshold were classified as higher risk areas, and the remaining ones were classified as low risk areas. This approach ensured that a similar kind of classification criterion was used during both model validation and spatial prediction processes.

Chapter 4. Results

4.1 Groundwater geochemistry

Samples were analyzed for five main types of anions, namely, nitrate (NO_3^-), chloride (Cl^-), sulfate (SO_4^{2-}), bromide (Br^-), and fluoride (F^-). The variation of concentrations of major anions across the study area as follows. (Table 4.1, Fig. 4.1). All groundwater geochemical data from this study are being made available in the Kansas Geological Survey AWQUA database.

NO_3^- was the most common contaminant in our private well water samples, with a significant variability over the study area, reflecting the influence of different types of factors and processes associated with NO_3^- contamination. NO_3^- concentrations ranged from below detection (<0.09 mg/L as N) to 55.4 mg/L as N, with a mean and median of 11.9 and 9.6 mg/L as N, respectively. Both the lower (5.1 mg/L as N) and upper (16.9 mg/L as N) quartile values of NO_3^- concentrations were well above 3 mg/L as N, which is the natural background level of the aquifer (Townsend & Young, 1995). Overall, the distribution of NO_3^- concentration was more positively skewed with several high-concentration outliers. When comparing county-wise, the average NO_3^- concentration in all counties exceed the EPA limit, except in Reno (9.82 ± 7.66 mg/L) & Kingman (9.95 ± 7.36 mg/L) counties in the eastern part of the aquifer. The highest mean NO_3^- concentration is reported from Edward County (17.07 ± 10.73 mg/L) in the west. The variability of the county-wise calculated mean values for NO_3^- is relatively limited compared to chloride and sulfate.

Cl^- concentration shows a considerable variation within those private well water samples (Figure 4.1), with values ranging from 3.0 mg/L to 1437.1 mg/L. We observed the highest mean Cl^- concentrations in the eastern portion of the aquifer from the wells sampled within Reno (171.44 ± 253.63 mg/L) and Rice (134.01 ± 89.48 mg/L) counties. Meanwhile, Kiowa County in

the southwest corner of the study area shows the lowest mean Cl^- concentration in the analyzed samples (14.53 ± 6.37 mg/L), showing a clear geographical separation from the areas with the highest mean Cl^- concentration.

County	F^- (mg/L)		Br^- (mg/L)		NO_3^- (mg/L)		Cl^- (mg/L)		SO_4^{2-} (mg/L)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Barton	0.41	0.13	0.18	0.13	13.31	13.79	117.13	96.99	75.16	87.04
Edwards	0.49	0.33	0.12	0.08	17.07	10.73	41.44	22.94	69.42	101.24
Kingman	0.29	0.14	0.09	0.07	9.95	7.36	52.82	64.16	124.61	384.57
Kiowa	0.26	0.04	0.05	0.01	10.06	6.13	14.53	6.37	14.60	3.62
Pawnee	0.49	0.23	0.22	0.15	14.13	8.70	92.50	71.66	66.63	85.63
Pratt	0.26	0.06	0.08	0.06	10.85	10.93	61.03	61.06	18.88	7.51
Reno	0.30	0.19	0.08	0.11	9.82	7.66	171.44	253.63	37.37	27.62
Rice	0.43	0.24	0.21	0.21	11.50	6.26	134.01	89.48	100.82	87.07
Stafford	0.33	0.08	0.10	0.05	11.55	8.78	90.81	201.22	33.75	27.37
Overall	0.36	0.19	0.13	0.12	11.97	9.47	95.13	146.99	55.35	117.99

Table 4.1: Summarized statistics of the anion data within the study area

The determined concentration values of SO_4^{2-} also show a considerable variability with a standard deviation of 117.99, with several samples consisting of elevated SO_4^{2-} concentrations above 100 mg/L. The highest mean SO_4^{2-} value was observed from Kingman County (124.61 ± 384.57 mg/L) in the southeast corner of the study area. This was followed by two counties in the northern part of the study area, Rice (100.82 ± 87.07 mg/L) and Barton (75.16 ± 87.04 mg/L) \ counties. But in the southern portion of the aquifer, two neighboring counties, Kiowa (14.60 ± 3.62 mg/L) and Pratt (18.88 ± 7.51 mg/L), showed a significantly lower SO_4^{2-} concentration.

In contrast to those dominant anion types, F^- concentration was less variable. There are a few outliers that are with concentrations exceeding 1.0 mg/L, as it is visible in the boxplot (Figure 4.1). Br^- is also showing a significantly low concentration in our study area, showing trace-level concentrations with an overall mean of 0.13 ± 0.12 mg/L, and it remains the same for the county-wise data as well. Further, the box plot clearly shows that the Br^- data is clustered

within a very narrow range. There are a few outliers with elevated concentrations, but the concentration is still lower than 1 mg/L. Nitrite (NO_2^-) concentrations in these samples are beyond the detection limit of the instrument.

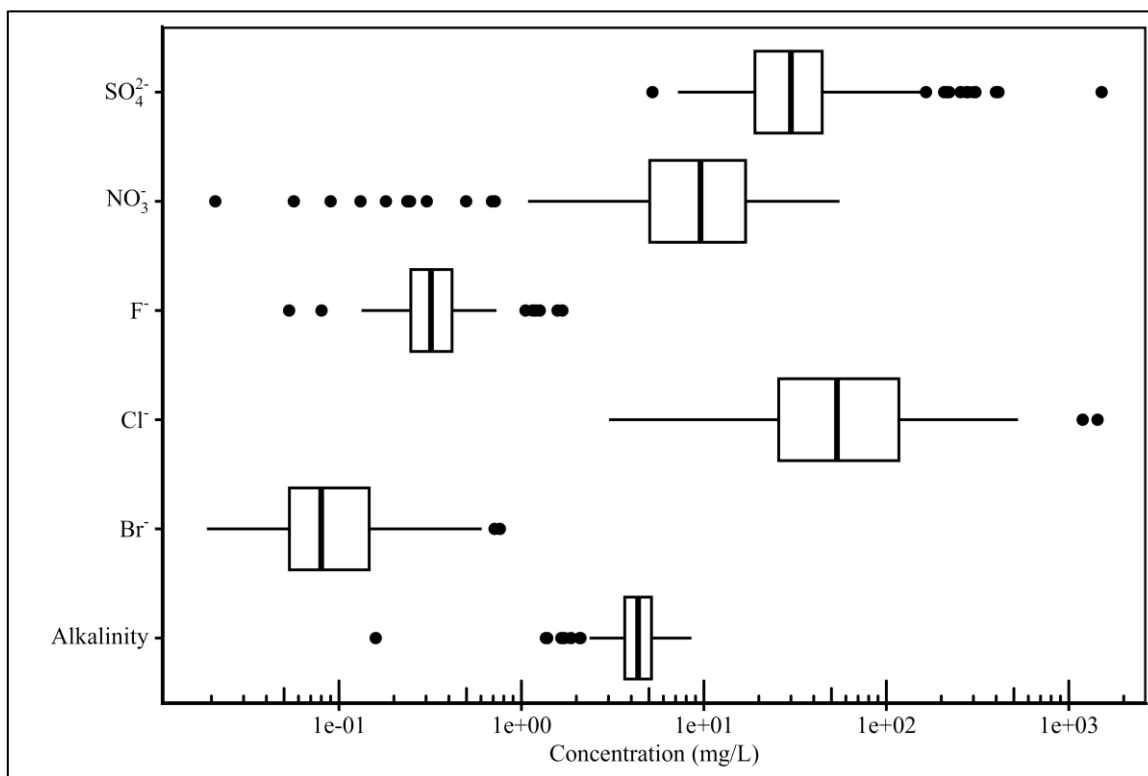


Figure 4.1: Box plot representing the variability of the anion data within the study area

Using the ion chromatography method, we were able to determine the concentrations of sodium (Na^+), potassium (K^+), ammonium (NH_4^+), magnesium (Mg^{2+}), calcium (Ca^{2+}), and strontium (Sr^{2+}). A summarized overview of the cation types that were determined in our geochemical analysis is shown in Table 4.2 and Figure 4.2.

Sodium (Na^+) has the second-highest overall mean among the cation types we determined. The calculated mean is 69.14 mg/L, but with a significantly high standard deviation (92.03 mg/L), which indicates a significant variation in our sampled wells. This is clearly observable with the box plot as well. Similar to the highest mean Cl^- values, we observed the highest mean Na^+ concentrations associated with Reno (124.95 ± 149.57 mg/L) and Rice (92.52 ± 57.51 mg/L) counties. Similarly, Kiowa County, located in the southwest corner of the study area, is showing the lowest average Na^+ value (14.72 ± 3.14 mg/L). This county-wise relationship evidenced that both sodium and chloride may result from a common process.

County	Na^+ (mg/L)		K^+ (mg/L)		Mg^{2+} (mg/L)		Ca^{2+} (mg/L)		Sr^{2+} (mg/L)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Barton	85.83	63.29	6.71	6.35	110.74	46.42	15.39	10.64	3.43	2.06
Edwards	46.20	31.23	4.13	1.14	104.91	33.56	14.35	8.80	2.64	2.42
Kingman	57.67	93.38	1.79	1.05	76.57	52.14	16.01	30.77	2.78	3.27
Kiowa	14.72	3.14	2.90	0.39	70.70	13.74	6.84	1.15	0.10	0.00
Pawnee	60.55	31.80	4.61	2.19	107.94	31.61	15.43	12.59	3.16	1.36
Pratt	37.68	33.38	2.96	0.96	86.20	28.83	6.92	2.05	0.83	1.15
Reno	124.95	149.57	3.21	2.27	86.50	33.17	9.54	5.30	2.61	0.96
Rice	92.52	57.51	3.28	1.07	118.61	37.15	18.96	10.48	3.97	1.54
Stafford	64.08	122.83	5.34	4.09	90.98	20.51	8.59	3.21	2.55	0.67
Overall	69.14	92.03	4.20	3.43	95.94	35.42	11.95	11.48	2.56	1.83

Table 4.2: Summarized statistics of the cation data within the study area

We observed the highest mean and median associated with Mg^{2+} in this cation analysis. In the county-wise comparison, Rice County is showing the highest mean (118.61 ± 37.15 mg/L), closely followed by Barton County (110.74 ± 46.42 mg/L) and Pawnee County (107.94 ± 31.61 mg/L). All these are neighboring counties that represent the northern portion of the Great Bend Prairie Aquifer. Similar to Na^+ , the lowest concentration was once again recorded from the wells' samples within Kiowa (70.70 ± 13.74 mg/L) in the southwest corner of the aquifer. In

comparison with Na^+ , the Mg^{2+} box plot is clustered within a narrow region with few higher outliers.

Calcium (Ca^{2+}) levels in the private wells within our study area are showing moderate variability across different counties, with an overall mean of 11.95 ± 11.48 mg/L. The highest and lowest mean calcium values are recorded from Rice (18.96 ± 10.48 mg/L) and Pratt counties, respectively, and they do not show any spatial correlation in positioning in two different corners of the aquifer.

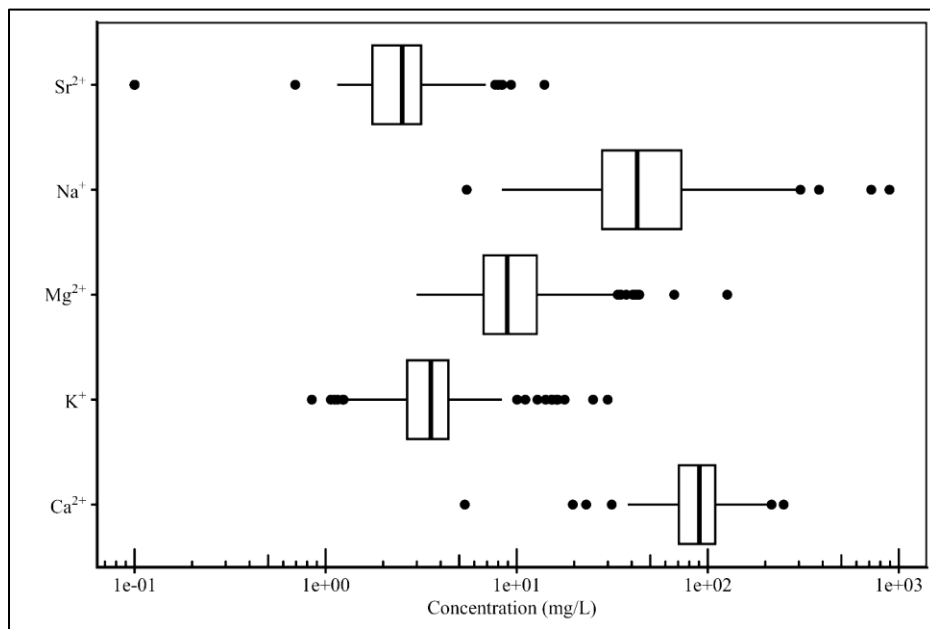


Figure 4.2: Box plot representing the variability of the cation data within the study area

Compared to those dominant cation types, both K^+ and Sr^{2+} show significantly lower mean concentration values overall as well as at the county level. The overall mean concentration of these two cation types is 4.20 ± 3.43 mg/L and 2.56 ± 1.83 mg/L, respectively. The highest mean potassium values are recorded from the wells within Barton County (6.71 ± 6.35 mg/L), and the lowest value is in Kingman County (1.79 ± 1.05 mg/L). Rice is the county with the highest mean strontium concentration (3.97 ± 1.54 mg/L), closely followed by all the other counties except Pratt and Kiowa counties, which have wells with strontium levels below the

detection levels of the instrument. Here, the statistic for ammonium (NH_4^+) is not included as the contamination levels are beyond the detection limit of the instrument in the majority of samples.

Piper diagrams that were prepared for the hydrogeochemical facies identification showed a hydrogeochemical clustering based on both similarity in ionic composition and their relative spatial position within the study area. Based on our grouping, we can identify four main subgroups associated, indicating that hydrogeochemical variation is systemic, which is controlled by the regional hydrogeochemical patterns.

The first group consists of Barton, Stafford, and Rice counties, geographically positioned in the central and northern portion of the study area, mainly consisting of Ca-HCO_3 type water (Fig. 4.4 a). The majority of samples are showing a clustering towards the Ca^{2+} apex in the cation triangle, with a relatively smaller contribution from the smaller contributions from Mg^{2+} and limited $\text{Na}^+ + \text{K}^+$ concentrations. Within the anion triangle, samples are mostly concentrated around HCO_3^- apex with a deviation towards Cl^- .

The second group consists of Reno and Kingman counties, representing the southeastern region of the aquifer. The cation plot indicates a spread between the Ca^{2+} apex and the $\text{Na}^+ + \text{K}^+$ apex, with a more dominance toward the first one. Anion distribution also shows a spread between HCO_3^- and Cl^- apexes (Fig. 4 .3 b). Several samples are showing a significant Cl^- enrichment, evidencing saline water mixing in some part of this region.

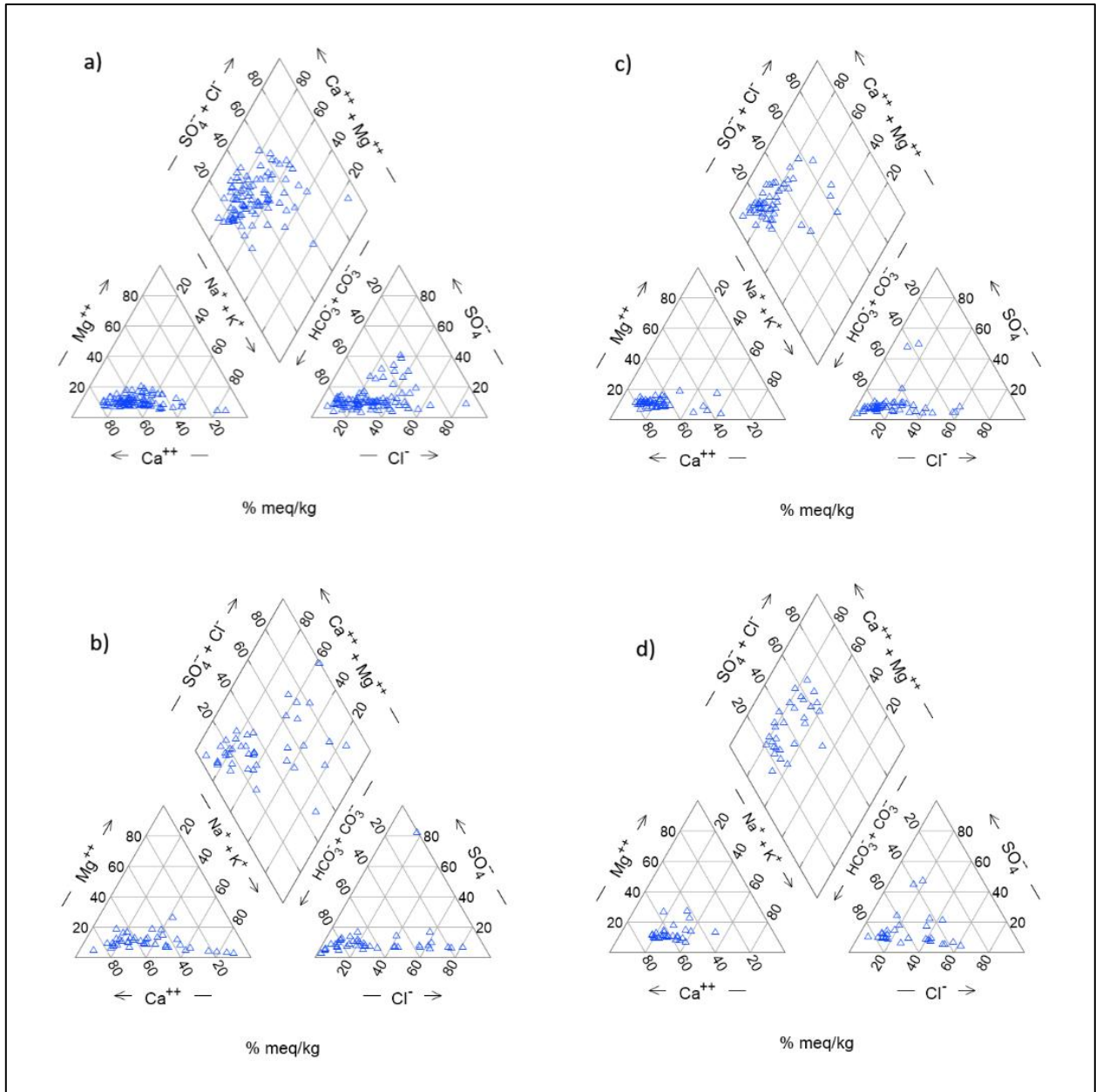


Figure 4.3: Piper diagrams showing grouped hydrogeochemical facies of groundwater in the study area: (a) Barton–Stafford–Rice (b) Reno–Kingman (c) Pratt–Kiowa–Edwards (Ca–HCO₃ dominant), and (d) Pawnee

The third group consists of the Pratt, Kiowa, and Edwards counties from the southern and southwestern part of the study area. They are showing a strong clustering towards Ca²⁺ and HCO₃⁻ apexes (Fig. 4.3 c), indicating that water is more Ca–HCO₃ type. Compared to the

previous two groups, these counties are showing a relatively smaller amount of impact from saline water mixing processes.

Pawnee County, in the western part of the aquifer, shows an intermediate position between the Ca^{2+} apex and the $\text{Na}^+ + \text{K}^+$ apex. However, the anion triangle is showing a mixed stage between all three apexes with slight dominance towards $\text{HCO}_3^- + \text{CO}_3^{2-}$.

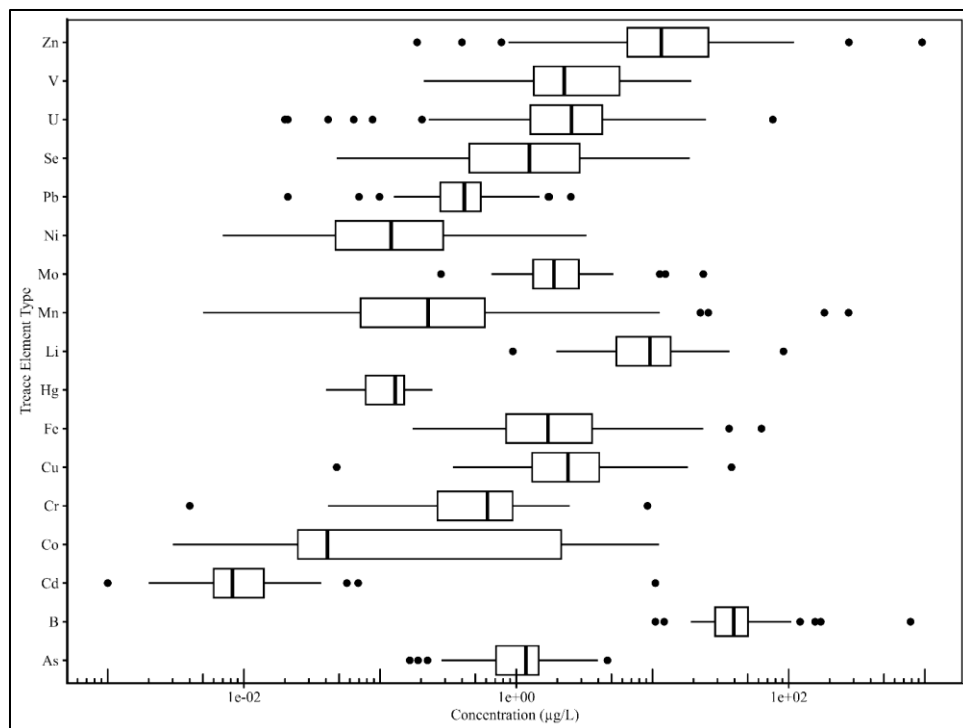


Figure 4.4: Distribution of trace element concentrations in groundwater samples from the study area.

Trace element concentrations were relatively low but variable (Fig. 4.4). For some of these trace elements, such as chromium (100 µg/L), copper (1300 µg/L), arsenic (10 µg/L), selenium (50 µg/L), mercury (2 µg/L), lead (15 µg/L), and uranium (30 µg/L), have MCLs defined by EPA. In our comparison of these measured contaminant values with the MCLs, we did not find any samples that exceeded the EPA levels, except one sample (10.496 µg/L) from Pawnee

County, which exceeded the MCL for cadmium. However, it can be identified as a localized occurrence, as none of the other samples show a similar level of high contamination.

Among the samples that we analyzed for PFAS, only one well from Kingman County produced groundwater with elevated PFAS concentration. All the other private well water samples contain no or an undetectable concentrations of PFAS compounds. The sample from well 25-KM-07 in Kingman County contained detectable perfluorooctanoic acid (PFOA) and perfluorooctane sulfonamide (PFOSA). However, the PFOA level (1.12 ng/L) for groundwater from this well was below the maximum contaminant level defined by the EPA (4 ng/L). Compared to that, an elevated concentration of PFOSA (8.40 ng/L) was determined within this sample. However, there is no defined MCL level for PFOSA to assess whether this is bad for humans or animals. All the other private well water samples that we analyzed are free of any detectable amount of PFAS, indicating that there is no widespread PFAS contamination issue associated with private wells in south-central Kansas.

4.2 Stable isotope composition within private well water

For most of the groundwater samples, $\delta^2\text{H-H}_2\text{O}$ ranged from -50 to -30‰ and $\delta^{18}\text{O-H}_2\text{O}$ ranged between about -7.5 and -5.5‰ and were largely aligned with the global and local meteoric water lines (Fig. 4.4). However, six samples had isotope ratios that were substantially lower, with $\delta^2\text{H-H}_2\text{O}$ between -70 to -55‰ and $\delta^{18}\text{O-H}_2\text{O}$ between -10 and -8‰. Those samples all came from wells located along the northern margin of the Great Bend Prairie Aquifer near the Arkansas River. Additionally, two samples had relatively high $\delta^{18}\text{O-H}_2\text{O}$ ratios and plotted considerably to the right of the local meteoric water line.

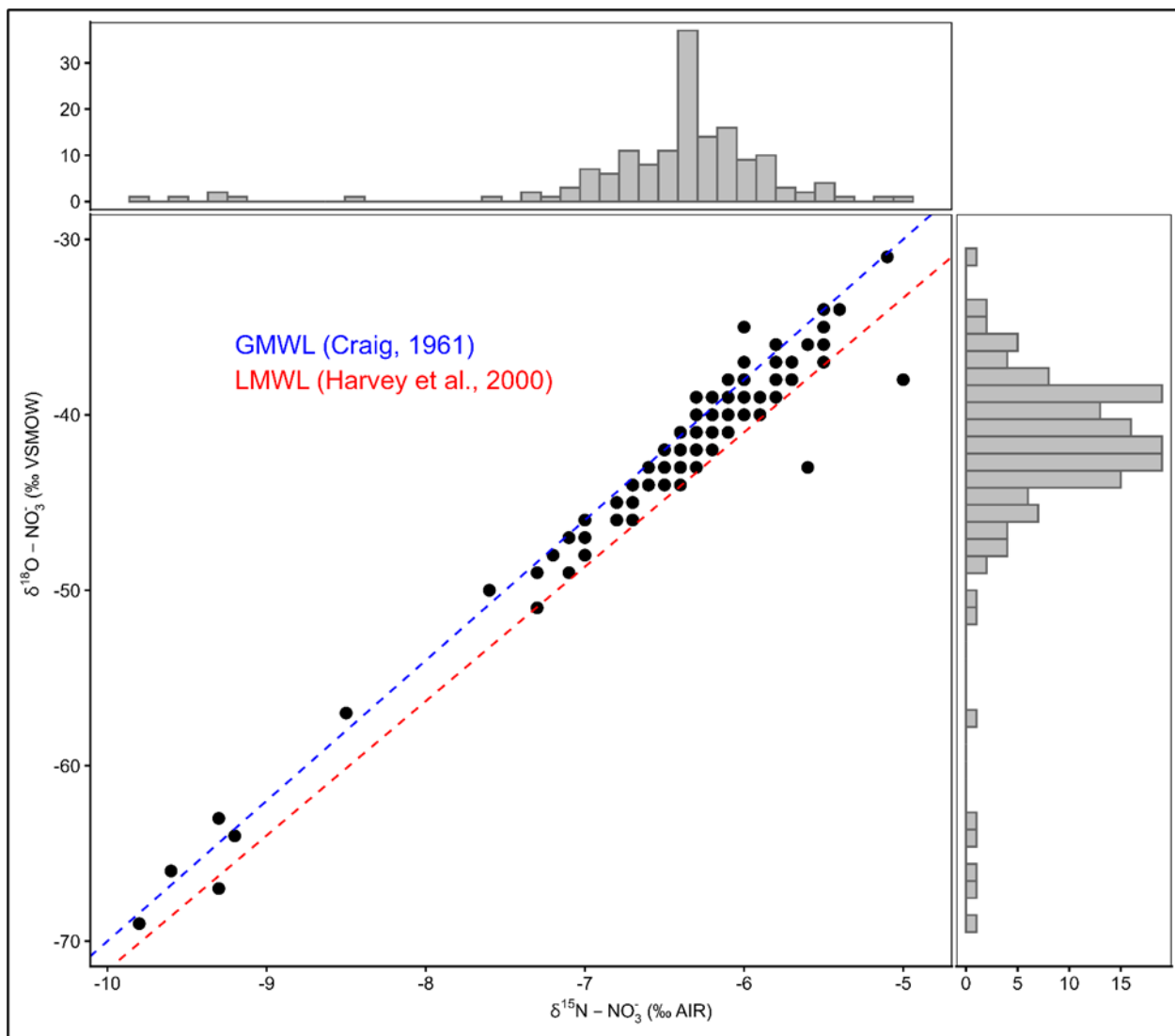


Figure 4.5: The stable isotope composition plot $\delta^2\text{H}$ and $\delta^{18}\text{O}$ in private well water (The Global Meteoric Water Line (GMWL; Craig, 1961) and Local Meteoric Water Line (LMWL; Harvey & Welker, 2000) are shown for reference.)

The isotopic composition of NO_3^- in our samples falls within or near the range expected for NO_3^- fertilizer (Fig. 4.6) and just above the range of values measured previously from observation wells in the Great Bend Prairie Aquifer by (Lane et al., 2020), which were largely consistent with ammonium-based fertilizer, perhaps with some inputs of NO_3^- from animal waste.

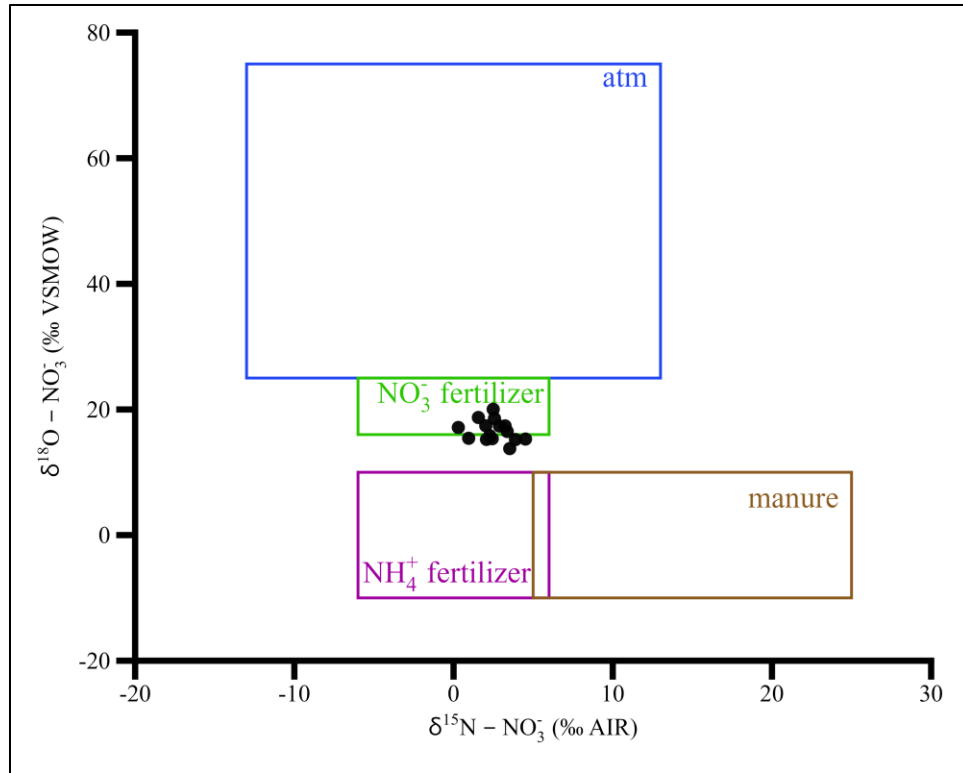


Figure 4.6: isotopic composition values of $\delta^{15}\text{N}-\text{NO}_3^-$ and $\delta^{18}\text{O}-\text{NO}_3^-$ in private well water samples with reference fields from (C.-Q. Liu et al., 2006) for potential nitrate sources, including atmospheric deposition, NO_3^- fertilizers, NH_4^+ fertilizers, and manure

4.3 NO_3^- distribution and relationships to other water quality parameters

NO_3^- concentrations vary widely across the study area (Fig. 4.7). 72 to 95% of the wells sampled in each county had NO_3^- concentration above 3 mg/L as N, which is considered to be the natural background level for the NO_3^- in groundwater (Townsend & Young, 1995). 31 to 48% of the wells in each county except Pawnee and Edwards had NO_3^- concentration above 10 mg/L as N. In Pawnee and Edwards counties, however, 62 and 79% of the wells, respectively, had NO_3^- concentrations above 10 mg/L as N (Fig. 4.7).

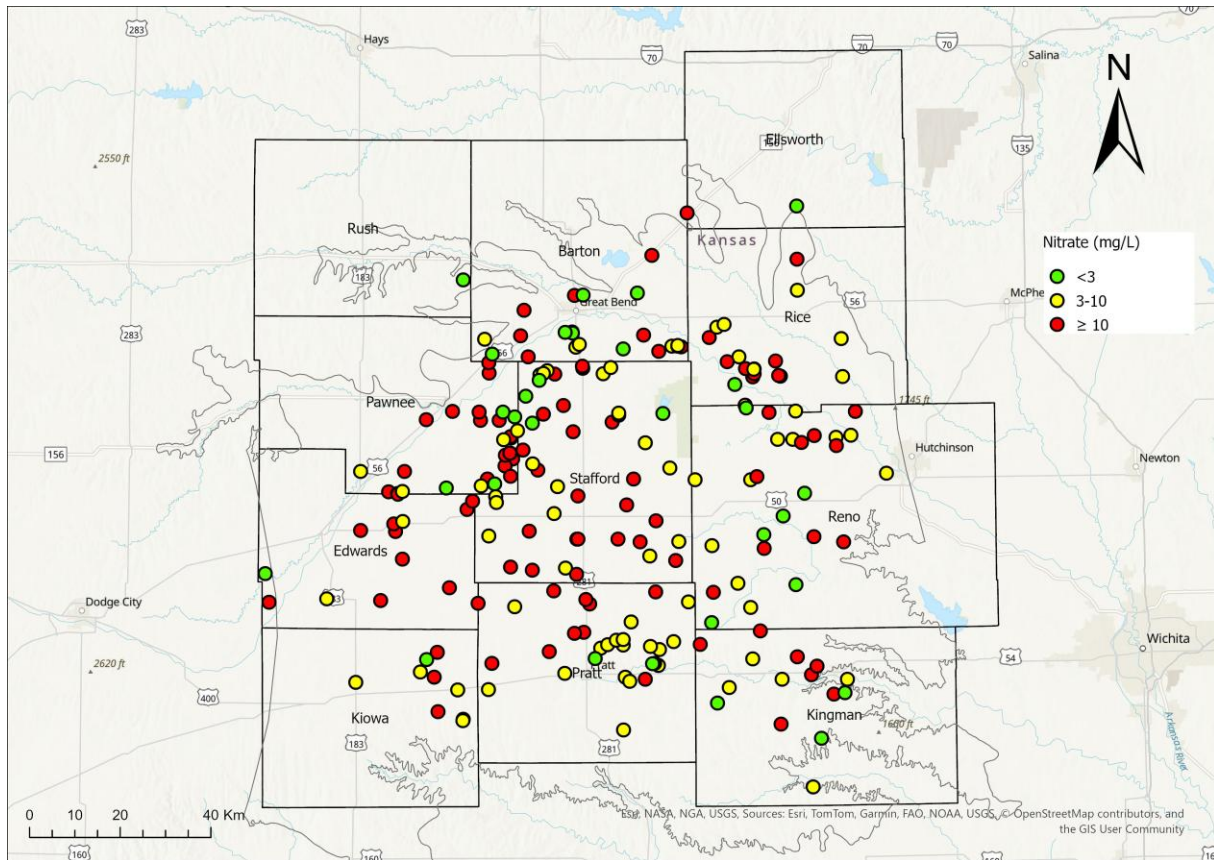


Figure 4.7: Map of NO_3^- contamination levels within the sampled wells

Table 4.3 summarizes the proportion of wells in each county that exceeded the natural background level (3mg/ L as N) (Townsend & Young, 1995) and EPA’s maximum contaminant level (10mg/L as N). Here, the percentage of wells exceeding the MCL within the entire study remains at a high value, like 46.5%. But it does not guarantee that all the remaining wells are safe, as 44.7 % of the wells have already exceeded the natural background levels of the aquifer, approaching the MCL.

When assessing these percentage values county-wise, the number of wells exceeding the MCL level reaches up to 78.6% and 62.1% within Edwards and Pawnee Counties, respectively, which are located in the western corner of the aquifer. In both Kiowa and Pratt counties in the

County	Percentage of wells (%)		
	<3 mg/L	3–10 mg/L	>10 mg/L
Barton	28.0	24.0	48.0
Edwards	7.1	14.3	78.6
Kingman	20.0	33.3	46.7
Kiowa	11.1	55.6	33.3
Pawnee	17.2	20.7	62.1
Pratt	6.9	62.1	31.0
Reno	22.2	40.7	37.0
Rice	5.0	50.0	45.0
Stafford	10.2	44.9	44.9
Overall	8.8	44.7	46.5

Table 4.3: Proportions of wells with NO₃⁻ concentrations below 3 mg/L as N, from 3-10 mg/L as N, and greater than 10 mg/L as N.

southern portion of the aquifer, all the sampled wells have already exceeded the background NO₃⁻ level of the aquifer. In contrast to that, around 20-30% of the wells sampled within Kingman, Barton, and Reno Counties have shown a NO₃⁻ level still below the natural background level. Overall, the spatial distribution and percentage-wise analysis of the NO₃⁻ levels within the study areas clearly evidence that this is a widespread issue with a large proportion of wells exceeding the natural background level of the aquifer and the EPA's maximum contaminant level.

4.4 Relationship between NO₃⁻ and hydrogeochemical parameters

NO₃⁻ shows a different kind of relationship with other dominant geochemical parameters like calcium (Ca²⁺), chloride (Cl⁻), dissolved oxygen (DO), hardness, magnesium (Mg²⁺), sodium (Na⁺), pH, and total dissolved solids (TDS) (Fig. 4 .8).

A very strong positive correlation can be observed between the measured Total Nitrogen (TN) level in groundwater and NO_3^- concentration, as both assess the dissolved nitrogen levels within water samples with differing specificity for nitrogen speciation.

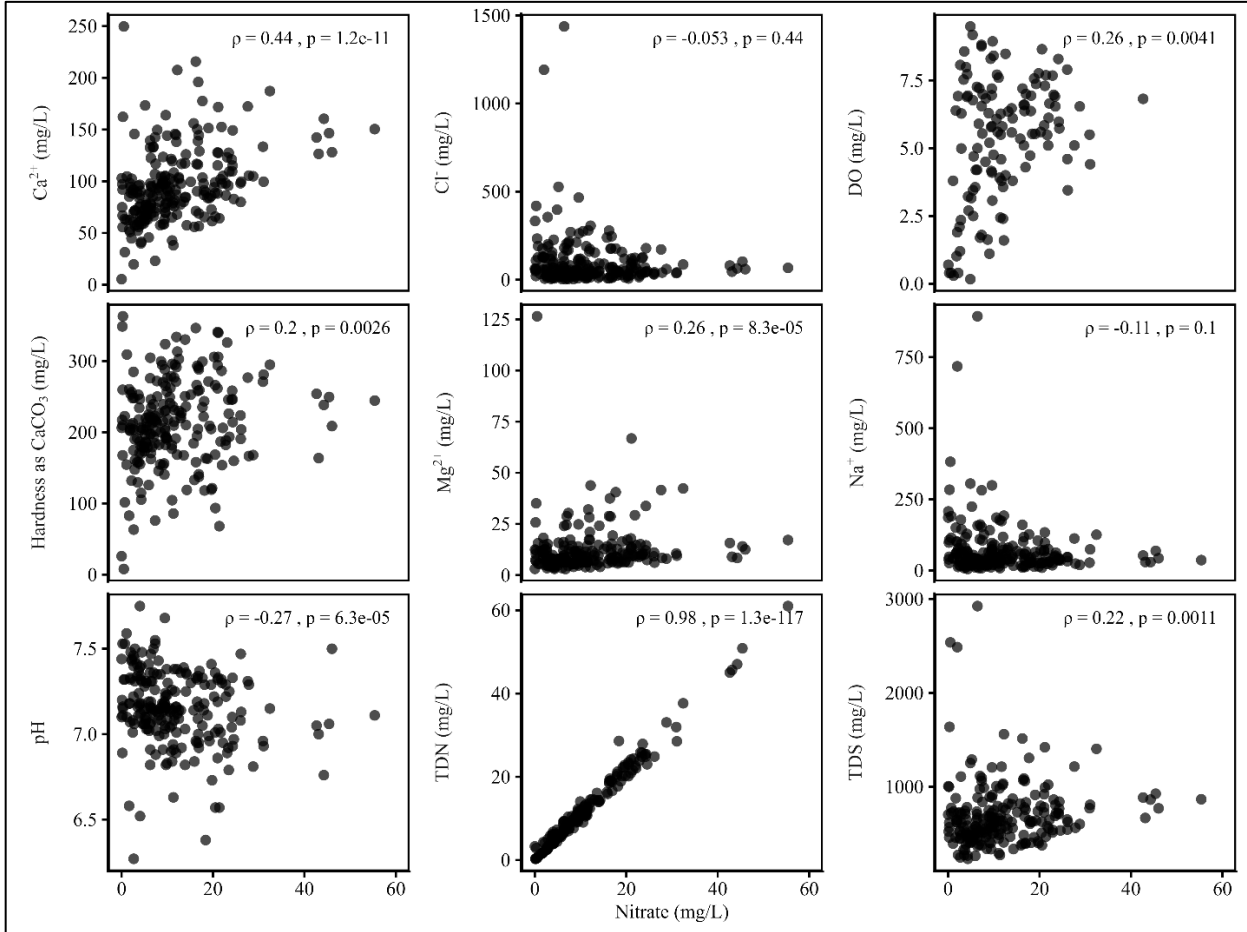


Figure 4.8: Relationships between NO_3^- concentration and selected geochemical parameters in groundwater samples.

We can observe a moderately positive correlation between NO_3^- and Ca^{2+} ($\rho = 0.44$, $p < 0.001$), indicating that wells with higher NO_3^- levels tend to have elevated Ca^{2+} concentration as well. Parameters like dissolved oxygen (DO) ($\rho = 0.26$, $p = 0.0041$), magnesium (Mg^{2+}) ($\rho = 0.26$, $p < 0.001$), carbonate hardness ($\rho = 0.20$, $p = 0.0026$), and total dissolved solids (TDS) ($\rho = 0.22$, $p = 0.0011$) had a weak positive correlation with NO_3^- . However, all those correlations were statistically significant.

A statistically significant, moderately negative correlation is visible between NO_3^- concentration and pH values, such that higher NO_3^- concentrations tend to occur in samples with lower pH.

In contrast, the most dominant anion type within these anion samples, Cl^- , does not show any significant correlation with NO_3^- ($\rho = -0.053$, $p = 0.44$), consistent with a lack of common source or geochemical process. Similarly, Na^+ concentration lacks a relationship with NO_3^- , with ρ and p values that are not statistically significant.

In summary, parameters like Ca^{2+} , Mg^{2+} , DO level, carbonate hardness, and TDS show a moderate or weak positive correlation with NO_3^- concentration, and on the other hand, pH is the only thing that shows a statistically significant negative correlation out of these.

4.5 Relationship of NO_3^- concentration to land use and land cover conditions

Below, we have plotted (Fig 4.9) the relationship of different land use/cover and crop types with NO_3^- s, and to assess their patterns statistically, we have calculated Spearman's rho value and p-value for each of the plots.

Here, we observed a moderate positive correlation with corn cultivated areas ($\rho = 0.37$, $p < 0.001$), and a weak positive correlation with the land use conditions like cultivated land ($\rho = 0.18$, $p = 0.007$), developed areas ($\rho = 0.19$, $p = 0.005$), and soybean cultivation ($\rho = 0.22$, $p < 0.001$).

A statistically significant, weak negative relationship can be observed associated with herbaceous land ($\rho = -0.18$, $p = 0.0092$). Even though the rho value is negative (-0.12), the relationship between NO_3^- and forest area cannot be considered statistically significant due to its p-value (0.074). Other dominant cultivation types, like sorghum cultivation ($\rho = 0.014$, $p = 0.84$) or winter wheat cultivation ($\rho = 0.07$, $p = 0.31$), also do not show any meaningful correlations with the NO_3^- concentration.

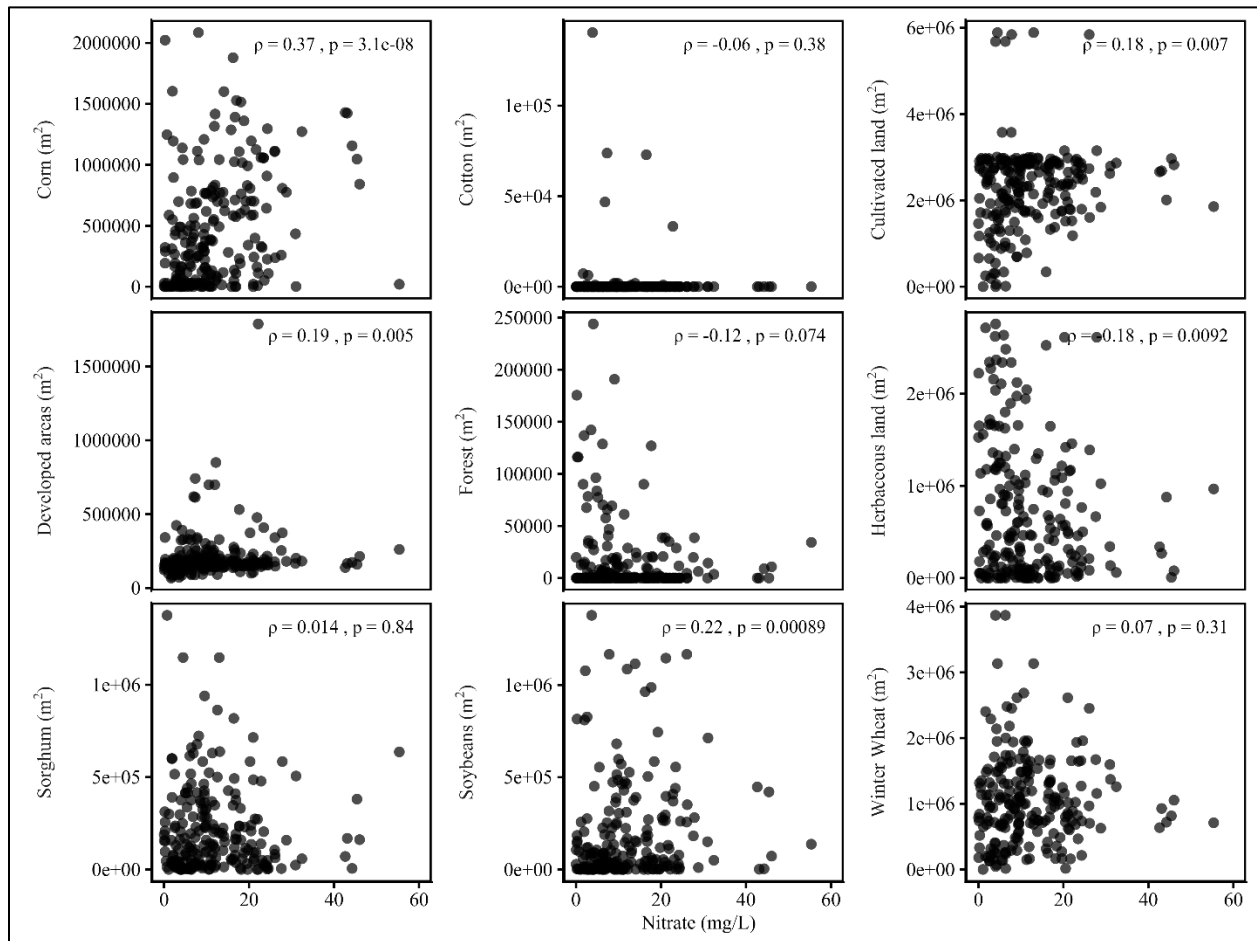


Figure 4.9: Relationships between NO_3^- concentration (mg/L) and land use/land cover variables (m²) within the study area.

4.6 Relationship of NO_3^- concentration to soil and hydrogeological properties

Hydrogeological conditions are very important in controlling the transport dynamics of the contaminants into the subsurface. Below, we have illustrated the relationship between NO_3^- concentration and selected hydrogeological parameters (Fig. 4.10). However, the Spearman's rank correlation analysis indicates that all these parameters have a weak correlation with NO_3^- .

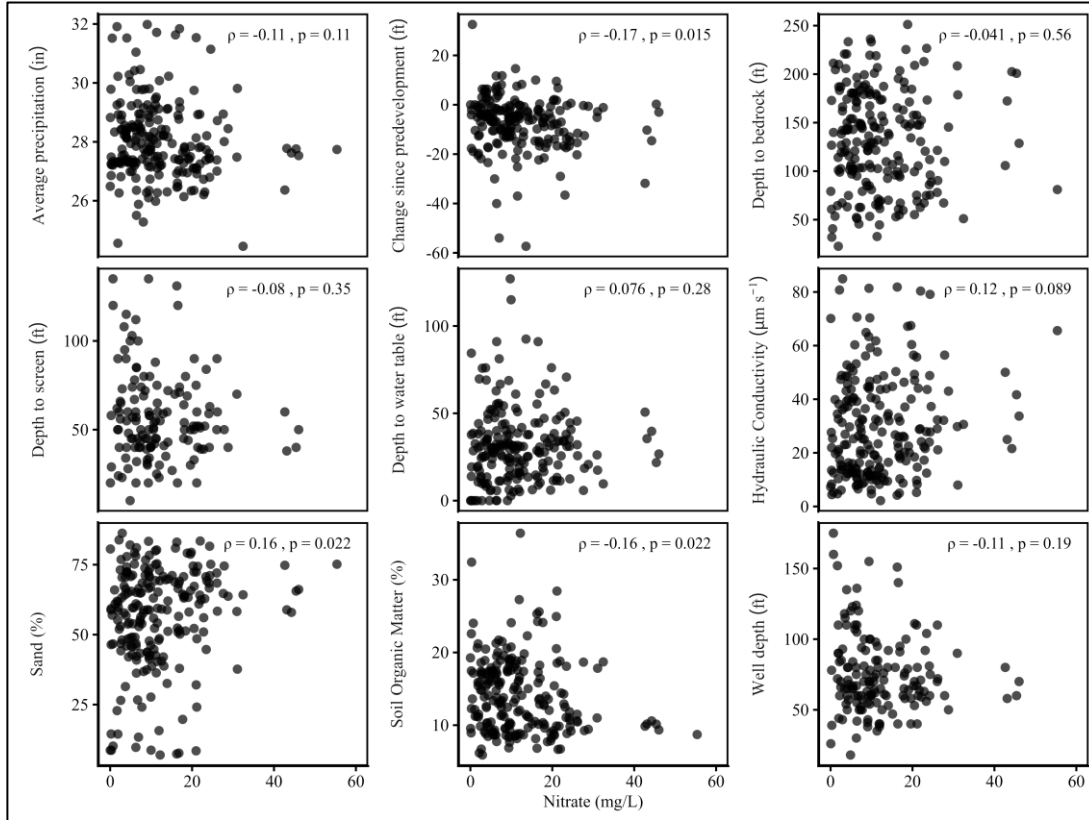


Figure 4.10: Relationships between NO_3^- concentration (mg/L) and hydrogeological and climatic parameters

A weak, but statistically significant and positive correlation can be observed between NO_3^- concentration and soil sand percentage ($\rho = 0.16$, $p = 0.022$). NO_3^- concentration was weakly related to hydraulic conductivity ($\rho = 0.12$, $p = 0.089$), soil organic matter content ($\rho = -0.16$, $p = 0.022$), change in water level since predevelopment ($\rho = -0.17$, $p = 0.015$), and well depth ($\rho = -0.11$, $p = 0.19$), though the relationships were insignificant. Even weaker relationship of NO_3^- concentration occur with depth to well screen ($\rho = -0.08$, $p = 0.35$), depth to

water table ($\rho = 0.076$, $p = 0.28$), depth to bedrock ($\rho = -0.041$, $p = 0.56$), and average precipitation ($\rho = -0.11$, $p = 0.11$).

As individual parameters, all of them are showing very weak correlations, and some of those relationships are not even statistically significant. However, combined effects of multiple parameters have the potential to influence NO_3^- concentrations.

4.7 Principal Component Analysis

Principal Component Analysis (PCA) was carried out to identify the factors that are contributing to the NO_3^- variability in groundwater by reducing the dimensionality of the available parameters and preserving the relationship between the key contributors. The first four principal components explained 75.6% of the total variance, capturing a considerable amount of trend within this dataset.

The first principal component (PC1) explained around 29% of the total variance and primarily represents the effect of the agricultural and land use conditions. Herbaceous lands and average precipitation have shown a strong negative loading, while cultivated land, water table elevation, corn cultivation, and distance from streams showed strong positive loadings.

The second principal component (PC2) explained around 23.3% of the total variance, and it is considerably represented by soil and hydrogeological conditions, which control the subsurface groundwater movements and the effects from microbial actions. Soil organic matter (SOM) content represents the strongest negative loading, while the hydraulic conductivity shows the strongest positive loading. Further, other variables such as distance from streams, water table elevation, saturated thickness, and NO_3^- also loaded positively on PC2.

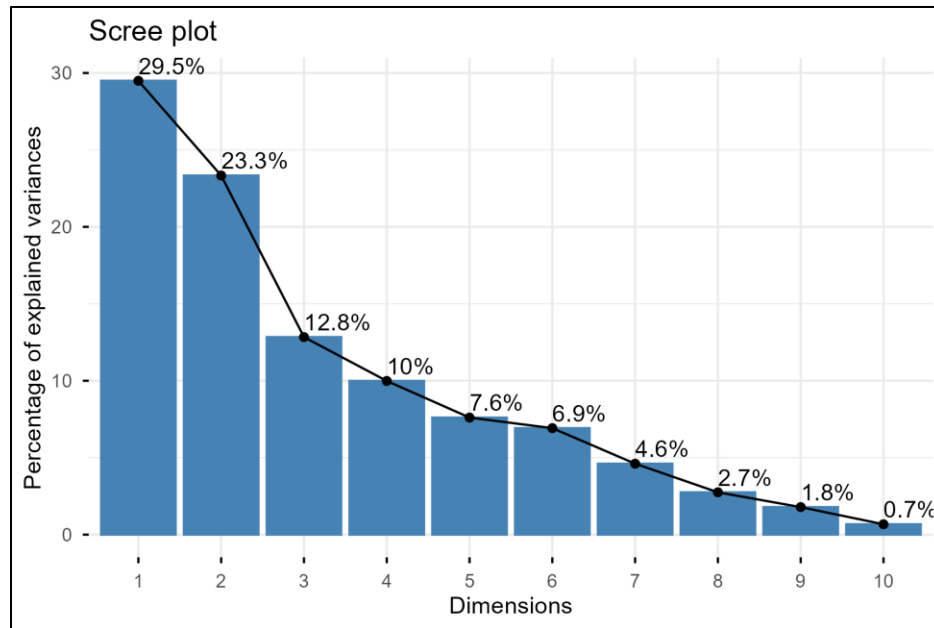


Figure 4.11: Scree plot showing variance explained by principal components, with PC1–PC4 accounting for ~75% of total variance.

12.8% of the total variance is explained by PC3, and it is dominated by the average annual precipitation, water table elevation, NO_3^- , and cultivated land. This component mainly represents the association of regional climate and groundwater availability. NO_3^- is positively correlated with average precipitation and negatively with water table elevation.

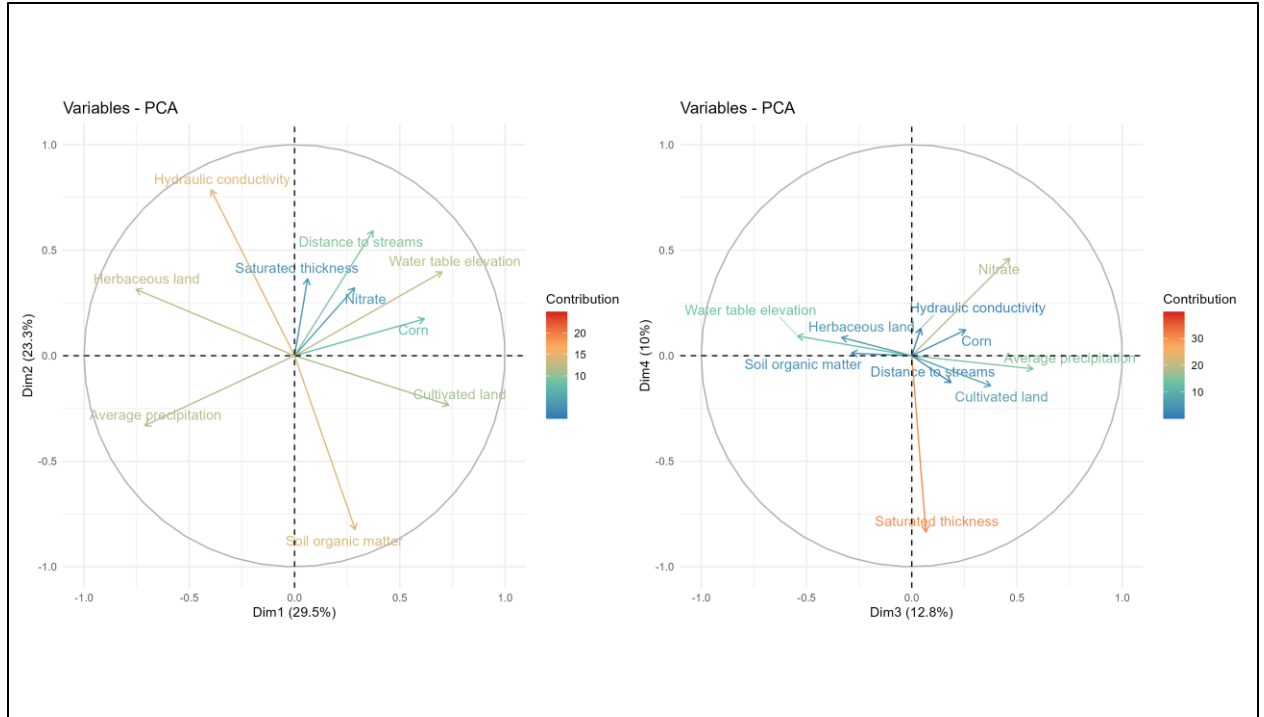


Figure 4.12: PCA loading plot showing nitrate aligned with corn cultivation along PC1 and hydrogeological controls (Ksat, depth to groundwater, average precipitation) along PC2.

The fourth principal component represents only 10% of the total variance, and it is mainly dominated by saturated thickness and NO_3^- , which are aligned in opposite directions. Saturated thickness is a measure of the extent of the aquifer, which provides evidence about the quantity of groundwater storage within the aquifer in the vicinity of each well.

4.8 Random forest analysis

Based on the automated process to tune the hyperparameters by testing different combinations, the final model was developed on 900 trees, $\text{mtry} = 3$, and $\text{min.node.size} = 12$ for optimized prediction ability. The Youden index-derived probability was 0.494, which was used as the threshold for classifying the wells.

The optimized model achieved a training AUC and out-of-bag (OOB) AUC of 0.998 and 0.749, respectively. The recorded independent test AUC was 0.702, indicating a good ability of the final model to identify the wells exceeding and not exceeding the EPA standard for NO_3^- . The model yielded an accuracy of 70.3%, with a sensitivity of 70.0%, specificity of 70.6%, precision of 67.7%, F1-score of 0.689, and balanced accuracy of 70.3%.

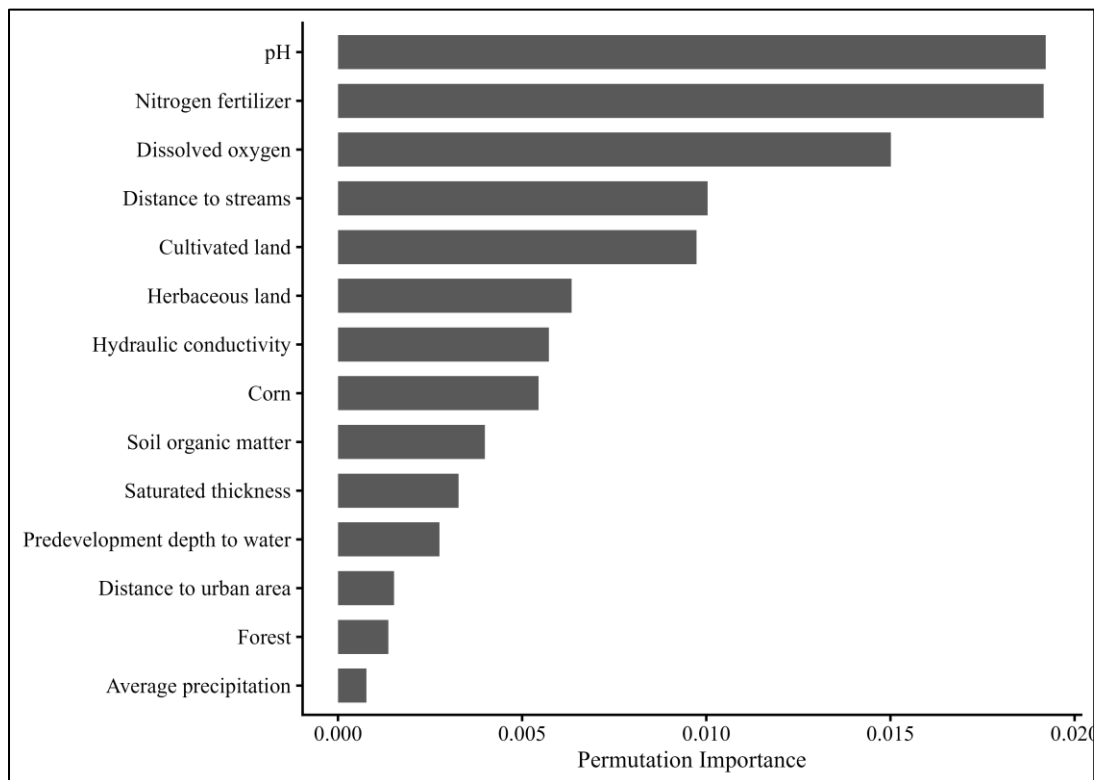


Figure 4.13: Variable importance in the random forest model

The variable importance plot (Fig. 4.13), which was constructed based on the permutation importance analysis, indicates pH, total nitrogen fertilizer application, and dissolved oxygen level are the top three most influential predictors in the model, followed by distance to stream and cultivated land. Variables such as herbaceous land cover, hydraulic conductivity, and corn cultivation are showing a moderate contribution. Other predictors such as soil organic matter, saturated thickness, predevelopment depth to water, distance to urban areas, forest cover, and average precipitation have shown lower importance compared to the top predictor variables.

Chapter 5. Discussion

5.1 Controls on NO_3^- concentrations in the study area

The relationship of NO_3^- in these groundwater samples with other geochemical and physical parameters helps to determine the potential control on the NO_3^- concentrations in the Great Bend Prairie Aquifer. Overall, these parameters represent the influence of land-use conditions, hydrogeological characteristics, and geochemical processes on NO_3^- variability in the study area.

5.1.1 Hydrogeochemical parameters

The positive relationship we observed between NO_3^- content and concentrations of Ca^{2+} and Mg^{2+} (Figure 4.8), alongside a similar relationship of NO_3^- concentration with carbonate hardness and TDS, suggests that processes that add NO_3^- contribute to increased concentrations of those cations and hardness. The High Plains Aquifer is identified as naturally rich with Ca and Mg rich water due to the dissolution of the carbonate-bearing minerals that are present within its sedimentary layers (Fader & Stullken, 1978; Hathaway et al., 1979). These dissolution processes can be enhanced by the NO_3^- -rich recharge water leaching through the soil and can enhance the release of Ca^{2+} and Mg^{2+} into groundwater.

When ammonium-based fertilizers are applied to the soil, the ammonium can be converted to NO_3^- by nitrification, which releases NO_3^- as well as protons (H^+) to the soil. When NO_3^- starts leaching to the subsurface, it moves alongside the H^+ , resulting in the dissolution of both carbonate and silicate rocks (Guo et al., 2015; Perrin et al., 2008), thus contributing to an increase in groundwater hardness and TDS. These are unlikely to be the only processes that govern these hydrogeochemical parameters, and therefore NO_3^- concentration is only moderate or weakly correlated to hardness, though the relationship is statistically significant.

Although chloride (Cl^-) and sodium (Na^+) are dominant ion types in many of our samples, neither correlates significantly with NO_3^- . This lack of a relationship likely reflects differences in source. Na^+ and Cl^- dominant groundwater enters the aquifer from the Permian bedrock flooring the aquifer (Whittemore, 1993) whereas NO_3^- inputs are dominantly from the surface. However, some studies have reported that Cl^- can also be introduced as an agricultural input and manure (Geilfus, 2019). But when coming to south-central Kansas, the overall Cl^- concentration is strongly governed by the natural evaporite dissolution, outweighing the impact from the agricultural inputs.

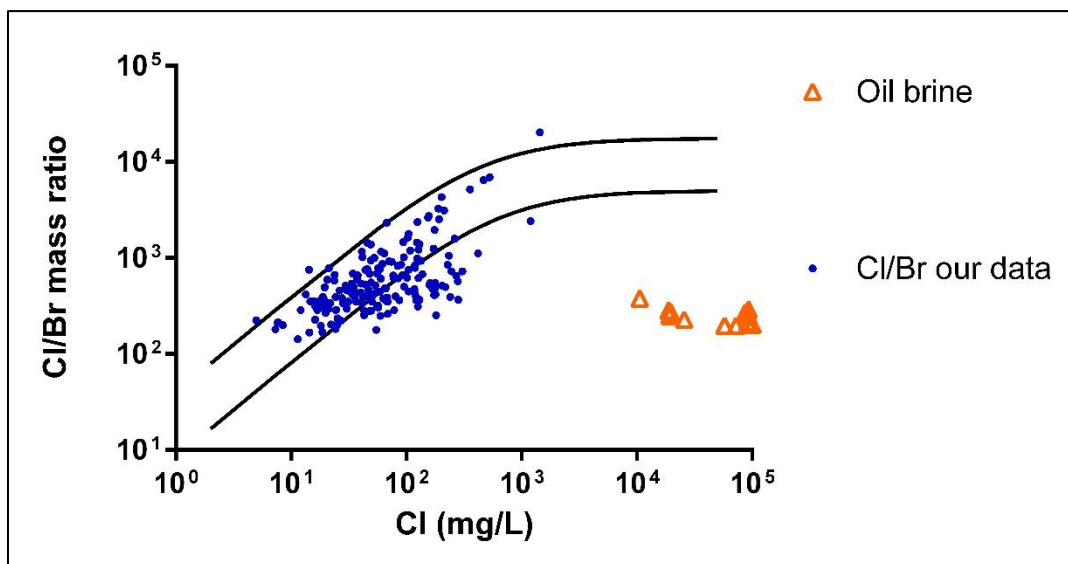


Figure 5.1: Cl^- concentration versus Cl/Br mass ratio for groundwater and oil brine samples (Whittemore, 1993)

This interpretation is consistent with results of a chloride-bromide mixing analysis (Fig. 5.1), which provides a better insight into the salinity and mixing processes in these groundwater samples. Most of the groundwater samples plot within a region of natural mixing defined by (Whittemore, 1993), evidencing a dominant mixing trend between freshwater and saline water, other than deriving from an independent salinity source. Bromide and chloride are considered to be the conservative constituents that are essentially unaffected by processes like precipitation,

redox, and adsorption processes associated with groundwater (Naily & Sudaryanto, 2018; Whittemore, 1993) and their concentrations need to be changed by mixing the two waters, and this narrow trajectory represent between the two end members. Here, the positioning of the samples within this band suggests that deriving saline water from a source with a low bromide concentration. In the natural halite precipitation processes with ancient seas, only a small amount of bromide is included, and in the dissolving process of these halite deposits within the Permian bedrocks, a high Cl/Br ratio can be expected increase in Cl^- concentration (Freeman, 2007; Whittemore, 1993). In contrast to that, Cl/Br ratios for oil brine samples plot outside this mixing trend, evidencing that they do not contribute significantly to the observed groundwater chemistry (Lane et al., 2020; Whittemore, 1993).

5.1.2 Microbial NO_3^- reduction

The relationship between NO_3^- and dissolved oxygen (DO) is evidence for microbial NO_3^- reduction within the aquifer. These two parameters show a weak but statistically significant positive correlation, indicating that low NO_3^- levels tend to occur in conditions where the dissolved oxygen level is lower. Where oxygen is available, microbial communities mainly use oxygen as their terminal electron acceptor (Kirk, 2024). However, where oxygen concentrations are low, microorganisms use alternative electron acceptors, such as NO_3^- . Microbial NO_3^- reduction pathways include denitrification and dissimilatory NO_3^- reduction to ammonia (DNRA). Both pathways consume NO_3^- and thus cause NO_3^- concentration to decrease. Therefore, the aquifer shows a higher potential for NO_3^- reduction in the zones with low DO levels compared to those with elevated DO concentrations.

One reason why the relationship between DO and NO_3^- was weak may be because the variables are only linked where DO levels are low. Where DO is high, NO_3^- reduction does not

occur. Under those conditions, NO_3^- reduction is limited and NO_3^- concentrations vary in response to other factors that are unrelated to DO concentration.

5.1.3 Land use and cover (LULC)

Positive relationships between NO_3^- concentration and LULC and crop types (Fig. 4.9) agree with our initial assumption that the fertilizer applied to these croplands has a direct impact on the NO_3^- contamination in the study area. The relationship of NO_3^- concentration to corn cultivation is stronger than other LULC and crop types, suggesting that it has a greater impact on the NO_3^- contamination. Our results agree with some of the previous studies, as some have reported on the extensive use of NO_3^- fertilizers on corn compared to other crop types (Patni & Culley, 1989), as well as on its strong association with NO_3^- contamination (Garcia et al., 2017). However, most of the other crop types show weak negative or statistically insignificant relationships with NO_3^- s, suggesting lower fertilizer application rates or more efficient nitrogen use can result in less nitrogen contamination.

In addition to crop cultivation, the positive correlation between NO_3^- concentration and developed areas indicates that the contribution from anthropogenic sources like septic systems and urban runoff may contribute to NO_3^- loading. However, low correlation suggests that their impact is much less compared to the contribution from agricultural activities.

Land cover types like forest show a negative correlation, but it is not statistically significant. Theoretically, they are not associated with any fertilizer inputs and thus have the potential to negatively correlate with NO_3^- concentration. However, subsurface flow can occur from cultivated areas regardless of the surface LULC boundaries, resulting in NO_3^- contamination in private wells close to the forest areas.

Although we observed multiple significant correlations of NO_3^- concentration with LULC and crop types, all of the relationships are either moderate or weak and not capable of fully explaining the contamination scenario observed in our study area. This result likely reflects that NO_3^- concentrations are governed by many factors and that a more advanced statistical analysis is needed to delineate these interconnected patterns among parameters.

5.1.4 Hydrogeological and environmental conditions

Hydrogeological and environmental conditions are known to serve as important controls on transport of contaminants from the surface. However, much like our analysis of LULC and crop types, pairwise correlations between NO_3^- and hydrogeological and environmental parameters revealed at best weak relationships.

Soils in the study area are very sandy (Fig. 5.2), which favors rapid drainage and thus high hydraulic conductivity values, a measure of the ease with which a fluid can flow through porous media. Both variables have positive correlations with NO_3^- concentration, as anticipated, though the relationships are weak.

County-level average precipitation has a weak negative correlation with NO_3^- concentration, consistent with dilution of the NO_3^- as precipitation increases (Dao et al., 2024). However, increasing precipitation can also contribute to the transport of surface contaminants to the subsurface (Sun et al., 2024). The weak correlation indicates the complexity associated with the effect of precipitation on NO_3^- contamination.

Well depth is considered to be an important parameter in groundwater studies (Lee & Murphy, 2020; Wheeler et al., 2015) as it is a deciding factor in terms of contaminant travel time through the subsurface. In a previous study related to this area, it was found that NO_3^- concentrations are relatively higher in shallower monitoring wells (Lane et al., 2020). However, NO_3^- concentrations do not show any statistically significant relationship with well depth for our analysis, causing a contradiction with previous findings.

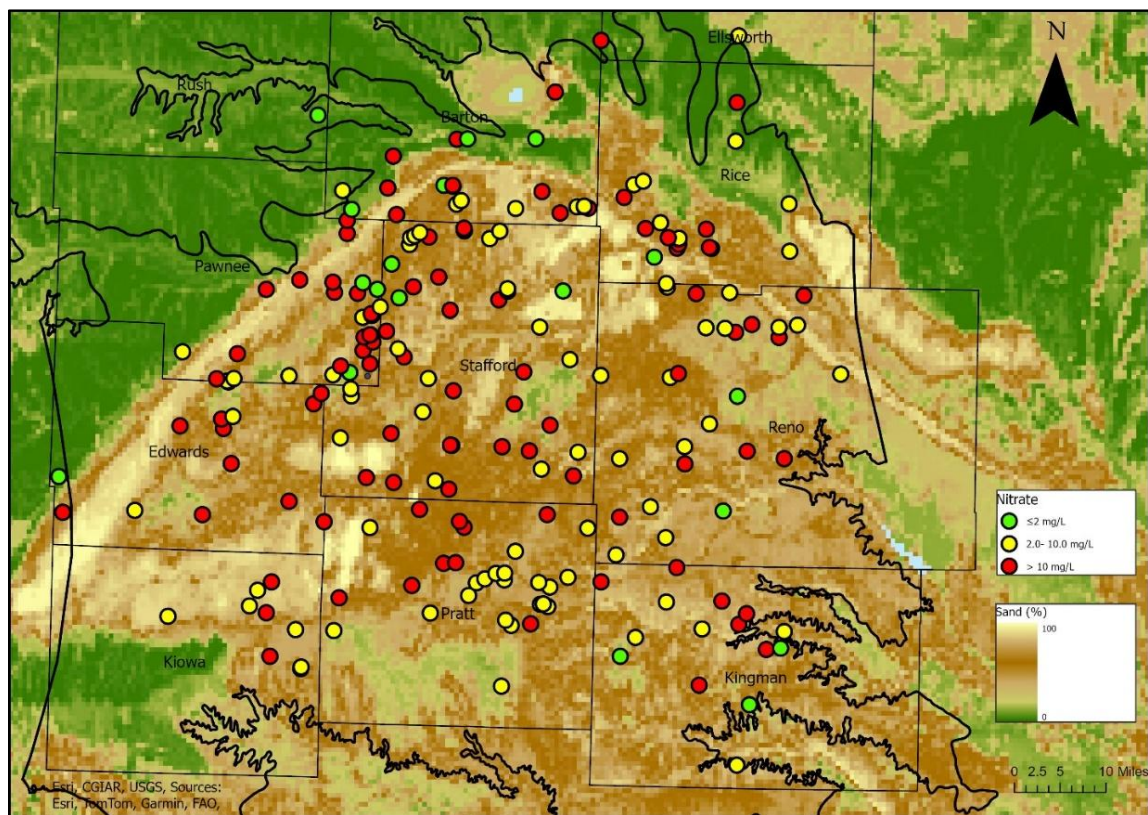


Figure 5.2: Percentage of sand within the soil (Data Source: UC Davies SoilWeb)

This difference may be controlled by a number of different factors. In the (Lane et al., 2020) study, well screen length was fixed around 5ft, with wells screened at the water table or at the aquifer base. For the private water wells we sampled, however, well design was variable, with well screens of varying lengths and even multiple well screens for some wells. All of these factors would serve to obscure relationships between depth and NO_3^- concentration. The depth to

the top of the screen varies from 10ft to 135 ft in the private wells we sampled, creating variability in terms of groundwater sources and subsurface residence times.

Apart from well depth, we examined the relationship between NO_3^- and other parameters that control the transport time of contaminants, such as depth to screen and depth to water table, as well. Again, none of these parameters correlated significantly with groundwater NO_3^- concentrations. The lack of strong correlations of transport parameters may reflect the superimposed effects of numerous factors controlling NO_3^- concentrations.

5.2 Principal component analysis (PCA)

Principal component analysis was carried out to identify the controls associated with groundwater NO_3^- contamination in the private wells. The first four principal components have been able to capture 75.6% of the total variance of the dataset, and this can be identified as a satisfactory level of achievement when comparing with similar kinds of groundwater quality studies (Li et al., 2020; Shrestha & Luo, 2017).

The positive correlation of NO_3^- and agricultural activities supports the assumption of NO_3^- contamination caused by intensive agricultural activities. Here, we included corn as a variable, as it showed a relatively significant correlation compared to the other crop types in our correlation analysis. On a global scale, corn can be identified as the largest consumer of synthetic N fertilizer, and it is responsible for consuming half of the N fertilizer used within the United States (Mirzaee & Mirzakhani Nafchi, 2025). Corn has shown a positive trend with NO_3^- s in PCA, evidencing that it is a main contributor of NO_3^- in groundwater. Apart from that, we included both cultivated lands and herbaceous land extent within a 1 km buffer zone around wells, and they have indicated positive and negative correlation, respectively, supporting our interpretation that groundwater NO_3^- concentration tends to be higher in wells surrounded by

cultivated lands, which are treated with fertilizer regularly. According to the calculations made based on published county-level fertilizer usage estimates (U.S. Geological Survey, 2023), counties in our study area fall within the top 25% of counties nationwide in terms of average fertilizer application per land surface area. Thus, it is a region with high N inputs.

NO_3^- positively correlate with hydraulic conductivity, which is the parameter responsible for facilitating the transport of surface contaminants into the subsurface.

The overlying sediment of the Great Bend Prairie Aquifer has a relatively higher hydraulic conductivity (Lane et al., 2020), increasing the vulnerability of the aquifer to get contaminated easily with NO_3^- . Water table elevation also positively correlates with groundwater NO_3^- concentration, evidencing a higher possibility of the aquifer being contaminated with NO_3^- when the water level is shallow (Townsend & Young, 1995), which reduces the advantage of having a longer travel distance through the vadose zone to limit the contaminations (Ascott et al., 2017). Saturated thickness was a dominant contribution to PC4 alongside NO_3^- , with each variable are negatively correlated, suggesting dilution effects associated with storage characteristics of the aquifer rather than any direct correlation with NO_3^- inputs. Some previous studies have also reported on the dilution effect of contaminants within the saturated zone compared to the unsaturated zone (Brouyère et al., 2004). Saturated thickness (Aquifer thickness) theoretically enhances the groundwater dilution in the dilution attenuation factor (DAF) models by being a variable that reduces the contaminant concentrations (Abrano Ic et al., 2001; Lasagna et al., 2013).

Distance to the streams positively correlates with NO_3^- , meaning wells that are located far away from the streams show higher NO_3^- concentrations. Previous studies have reported on the effect of the streams on groundwater, which creates a pathway for the movement of

contaminants. Depending on the difference between the groundwater table and stream level, either inflow or outflow happens from the aquifer, altering groundwater flow paths, residence times, and solute transport compared to those which are located far away from the streams, and these continuous interactions can cause changes in chemistry (Winter et al., 1998).

Soil organic matter negatively correlates with NO_3^- levels. When the soil organic matter level is high, it favors the development of anaerobic conditions, where microbial NO_3^- reduction could occur and lower NO_3^- levels. Here, organic matter serves as the electron donors to carry out the heterotrophic denitrification processes (Rivett et al., 2008).

The relationship between NO_3^- and precipitation appears to be complex. Precipitation was negatively correlated with NO_3^- along the first two principal components. But when coming to the third and fourth, it is positively correlated. This result suggests the contrasting impacts from precipitation, as a diluter but also transporter of NO_3^- , may occur in the study area, which exists on a precipitation gradient, with mean annual precipitation increasing eastward. Precipitation can percolate through soils and leach NO_3^- , ultimately transporting it to the aquifer and resulting in an accumulation of NO_3^- . However, if precipitation amounts are high enough, the NO_3^- entering the aquifer can be diluted. These results suggest that precipitation may be a contributor to elevated NO_3^- in the western portion of the study area where less precipitation occurs and a diluter of groundwater NO_3^- in the eastern portion of the study area, where precipitation rates are higher.

5.3 Random forest model

The random forest model has shown a satisfactory ability to identify the wells with contamination exceeding the EPA standards for drinking water by achieving a test AUC of 0.702. For this kind of model based on environmental data, a test AUC greater than 0.70

indicates acceptable discriminatory performance that is capable to differentiate high risk and low-risk wells with reasonable reliability despite the complexities of factors that govern NO_3^- contamination in a groundwater system. For the testing data, the overall classification accuracy of the model is 70.3%. The sensitivity and specificity are 70.0% and 70.6%, respectively, indicating that the model has the ability to identify high-risk and low-risk wells without being biased towards either class. The relatively smaller difference between OOB AUC (0.749) and independent test AUC (0.702) is good evidence for the model's capability to generalize to unseen data without any overfitting. The overall predictive performance of the random forest model reflects its ability to capture complex, non-linear relationships and interactions among groundwater chemistry, land use, agricultural practices, climate, and hydrogeologic factors when compared with conventional statistical approaches.

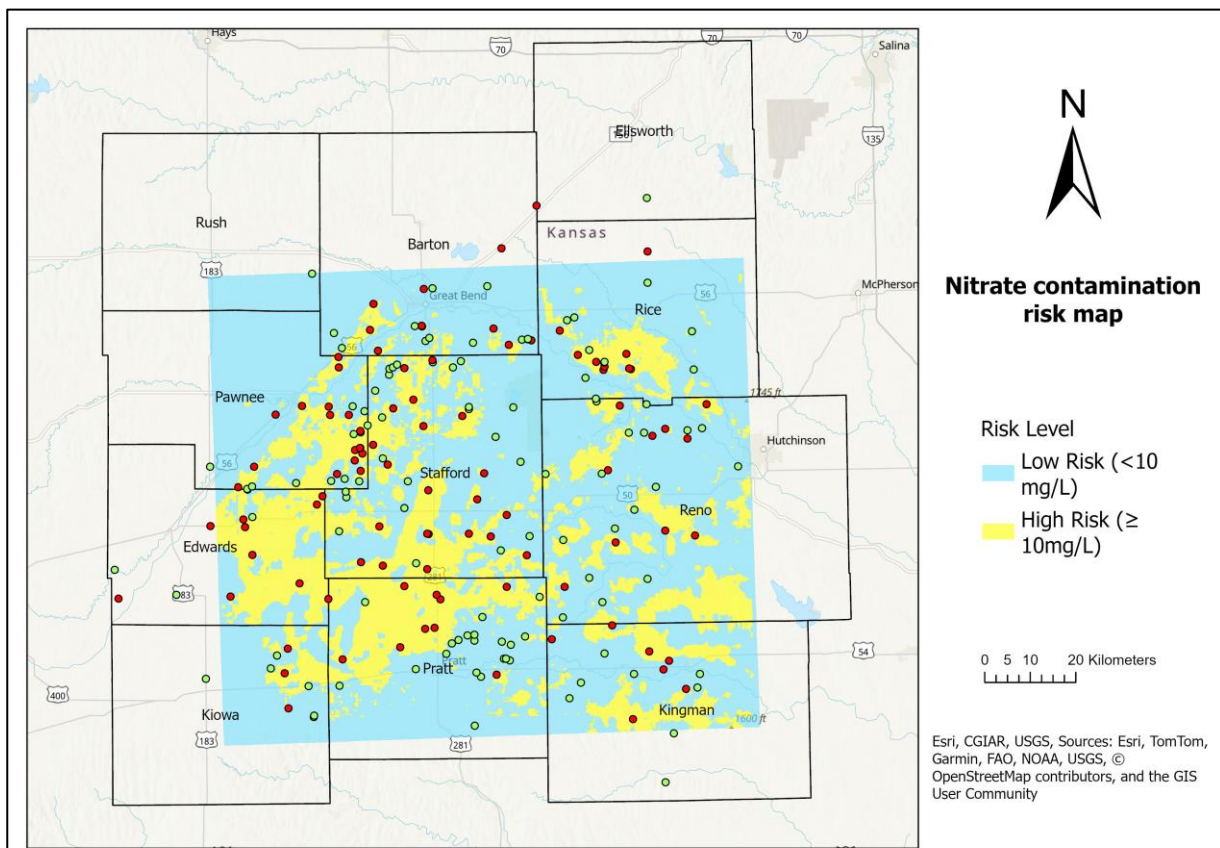


Figure 5.3: Nitrate contamination risk map developed using spatial prediction based on the random forest model

The random forest model was used to generate a spatial prediction map (Fig 5.3) to determine the low and high-risk areas. The comparison of the predicted risk map with the detected NO_3^- concentrations of the sampled wells indicates a good overall agreement. However, a small number of mismatches remain as the model cannot capture all the trends, especially the complexities such as localized changes in hydraulic conductivity, flow directions, and geologic materials, and different well techniques followed in the construction of each specific wells. This NO_3^- risk prediction map can be used as a base reference for future studies and groundwater management decision-making. This model and the prediction map can be improved further in the future with higher-resolution data, key variables missing in the current model, and also with more sampling points.

5.4 Comparison of NO_3^- contamination in private wells with other regions of the United States

The results of our study indicate that 46% of the sampled wells exceed the EPA's MCL, demonstrating that NO_3^- contamination is a widespread issue within the Great Bend Prairie Aquifer. In order to assess the nature of the severity of the contamination level within our study area, we compared our results with similar kinds of studies from other regions in the United States.

In a national-level study carried out by the United States Geological Survey (USGS), it was found that 25% of the groundwater wells tested in that study exceeded the MCL level for NO_3^- (Nolan et al., 1997). In comparison of our result data with the rest of the regions in the United States, it is clear that this is a significantly higher percentage, and similar kinds of high rates have been recorded only within a few of the regions in the country, which can also be

identified as areas with extensive agricultural activities. Even though some studies show a higher percentage of wells exceeding MCL, the issue is that their sample size is very small, limiting their representativeness, while increasing the uncertainty (Menzie et al., 2020).

In a study based on California's San Joaquin Valley region, the percentage of wells exceeding the MCL is over 40% in Stanislaus, Merced, Tulare, and Kings counties (Lockhart et al., 2013). In another study based on domestic and public wells in the Washington study, it found that the percentage of wells exceeding the MCL in Franklin and Douglas counties are 33% and 30%, respectively (Ryker & Jones, 1995). Leibbrand (1985) study that sampled 27 wells within the Big Sioux aquifer in South Dakota found that 40% of the wells are exceeding the MCL level of 10 mg/L as N. In south-central Wisconsin, which is a highly cultivated region, the percentage of private wells exceeding the MCL is between 20-30% (Wisconsin Groundwater Coordinating Council Wisconsin. Department of Natural Resources, 2021). Another study focused on the private wells located in the Umatilla and Morrow counties within the Columbia basin of central Oregon has found that 11 out of 25 wells (44%) are exceeding the MCL (S. L. Johnson et al., 1989). Most of these areas with relatively high percentages of wells with NO_3^- above the MCL are characterized by higher cultivated land percentage, higher fertilizer application rate, extensive irrigation, and permeable soil (Dubrovsky et al., 2010; Spalding & Exner, 1993).

However, in general, the majority of the county, aquifer, or state-level studies evidenced that the percentage of wells exceeding the MCL for NO_3^- is less than 10% in most of the studies, while reaching the percentage up to 20% in agricultural land use settings (Burow et al., 2010b). Thus, the proportion of wells with high NO_3^- that we observed places south-central Kansas among the worst cases of NO_3^- contamination, underscoring the severity of the problem.

5.5 Concluding remarks

This study is a comprehensive assessment of the groundwater quality of the private wells across the Great Bend Prairie aquifer in south-central Kansas. The overall results indicated that the dominant groundwater contamination issue associated with this aquifer is NO_3^- contamination, and over 46.5% of the sampled wells are above the US EPA MCL of 10 mg/L as N, while an additional 44.7% exceed the natural background level of the aquifer. When comparing with other regions in the United States, the Great Bend Prairie Aquifer falls among the most extreme cases of groundwater NO_3^- contamination.

Land use, land cover, and crop production play a significant role in the NO_3^- contamination, as it was evidenced by the results of the PCA. Among the different crop types cultivated in the area, corn cultivation had the strongest relationship to groundwater NO_3^- concentration, suggesting that corn production is one of the main contributors to NO_3^- contamination in the study area.

When assessing the variation of NO_3^- contamination with individual hydrogeological parameters such as depth to water table and hydraulic conductivity, as well as climatic factors such as average precipitation, they did not show any strong correlation. But when their combined impact was assessed through a multivariate analysis, those parameters showed a much more effective correlation, indicating the NO_3^- distribution is governed by complex interactions between surface inputs and subsurface transport processes.

These results indicate the urgent need for improved groundwater management practices to control the further deterioration of the water quality of the aquifer. Apart from that, this further highlights the importance of the introduction of federal or state-level regulation regarding private well construction and maintenance to improve community health.

Chapter 6. References

- Ablah, E., Marrow, M. W., Brown, J., & Honn, A. (2020). Analysis of Kansas Water Well Policies and Proposal of Nonpublic Household Water Well Recommendations. *Environmental Health Perspectives*, 128(2), 025001. <https://doi.org/10.1289/EHP5507>
- Abrano Ic, D., Johnson, P. C., Charbeneau, Randall. J., & Hemstreet, T. J. (2001). A Graphical Approach for Determining Dilution-Attenuation Factors: Basic Theory and Approach for Submerged Sources. *Groundwater Monitoring & Remediation*, 21(1), 115–124. <https://doi.org/10.1111/j.1745-6592.2001.tb00636.x>
- Adelana, S. M. A. (2005). Nitrate health effects. *Water Encyclopedia*, 4, 30–42.
- Anekwe, I. M. S., & Isa, Y. M. (2023). Bioremediation of acid mine drainage – Review. *Alexandria Engineering Journal*, 65, 1047–1075. <https://doi.org/10.1016/j.aej.2022.09.053>
- Arbogast, A. F. (1996). Stratigraphic evidence for late-Holocene aeolian sand mobilization and soil formation in south-central Kansas, USA. *Journal of Arid Environments*, 34(4), 403–414.
- Arif, M., Hussain, I., Hussain, J., Sharma, S., & Kumar, S. (2012). Fluoride in the Drinking Water of Nagaur Tehsil of Nagaur District, Rajasthan, India. *Bulletin of Environmental Contamination and Toxicology*, 88(6), 870–875. <https://doi.org/10.1007/s00128-012-0572-4>
- Ascott, M. J., Gooddy, D. C., Wang, L., Stuart, M. E., Lewis, M. A., Ward, R. S., & Binley, A. M. (2017). Global patterns of nitrate storage in the vadose zone. *Nature Communications*, 8(1), 1416. <https://doi.org/10.1038/s41467-017-01321-w>
- Atabati, A., Adab, H., Zolfaghari, G., & Nasrabadi, M. (2022). Modeling groundwater nitrate concentrations using spatial and non-spatial regression models in a semi-arid environment. *Water Science and Engineering*, 15(3), 218–227. <https://doi.org/10.1016/j.wse.2022.05.002>

- Baena-Moreno, F. M., Rodríguez-Galán, M., Arroyo-Torralvo, F., & Vilches, L. F. (2020). Low-Energy Method for Water-Mineral Recovery from Acid Mine Drainage Based on Membrane Technology: Evaluation of Inorganic Salts as Draw Solutions. *Environmental Science & Technology*, 54(17), 10936–10943. <https://doi.org/10.1021/acs.est.0c03392>
- Bakker, J. D. (2026). *Transformations*.
<https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/transformations/>
- Balke, K.-D., & Zhu, Y. (2008). Natural water purification and water management by artificial groundwater recharge. *Journal of Zhejiang University SCIENCE B*, 9(3), 221–226.
<https://doi.org/10.1631/jzus.B0710635>
- Böhlke, J.-K. (2002). Erratum: Groundwater recharge and agricultural contamination. *Hydrogeology Journal*, 10(3), 438–439. <https://doi.org/10.1007/s10040-002-0210-z>
- Bowen, K., Krishna, T., Backer, L., Hodgins, K., Waller, L. A., & Gribble, M. O. (2019). State-level policies concerning private wells in the United States. *Water Policy*, 21(2), 428–435.
<https://doi.org/10.2166/wp.2019.205>
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32.
<https://doi.org/10.1023/A:1010933404324>
- Brouyère, S., Dassargues, A., & Hallet, V. (2004). Migration of contaminants through the unsaturated zone overlying the Hesbaye chalky aquifer in Belgium: A field investigation. *Journal of Contaminant Hydrology*, 72(1–4), 135–164. <https://doi.org/10.1016/j.jconhyd.2003.10.009>
- Bruning-Fann, C. S., & Kaneene, J. B. (1993). The effects of nitrate, nitrite and N-nitroso compounds on human health: A review. *Veterinary and Human Toxicology*, 35(6), 521–538.

- Burow, K. R., Nolan, B. T., Rupert, M. G., & Dubrovsky, N. M. (2010). Nitrate in Groundwater of the United States, 1991–2003. *Environmental Science & Technology*, 44(13), 4988–4997.
<https://doi.org/10.1021/es100546y>
- Cabot, J. H., & Ross, E. G. (2023). Evaluating Prediction Model Performance. *Surgery*, 174(3), 723–726. <https://doi.org/10.1016/j.surg.2023.05.023>
- Cantlay, T., Bain, D. J., Dombrowski, C., Rykaceski, M., Bayto, M., Miller, K., & Stolz, J. F. (2025). Contamination of private water supplies after a well communication event (frac-out) in southwest Pennsylvania. *Scientific Reports*, 15(1), 32352. <https://doi.org/10.1038/s41598-025-16976-5>
- Chellam, S., & Clifford, D. A. (2002). Physical–Chemical Treatment of Groundwater Contaminated by Leachate from Surface Disposal of Uranium Tailings. *Journal of Environmental Engineering*, 128(10), 942–952. [https://doi.org/10.1061/\(ASCE\)0733-9372\(2002\)128:10\(942\)](https://doi.org/10.1061/(ASCE)0733-9372(2002)128:10(942))
- Chen, G.-F., Lai, C.-H., & Chen, W.-H. (2020). Principal Component Analysis and Mapping to Characterize the Emission of Volatile Organic Compounds in a Typical Petrochemical Industrial Park. *Aerosol and Air Quality Research*. <https://doi.org/10.4209/aaqr.2019.07.0365>
- Cherlinka, V., Gallay, M., & Dmytruk, Y. (2025). *Predictive Modeling*. https://uge-share.science.upjs.sk/webshared/uge_web_files/studium/ucebnice_skripta/Cherlinka_Gallay_Dmytruk_Predictive_Soil_Model_Textbook.pdf
- Cloutier, V., Lefebvre, R., Therrien, R., & Savard, M. M. (2008). Multivariate statistical analysis of geochemical data as indicative of the hydrogeochemical evolution of groundwater in a sedimentary rock aquifer system. *Journal of Hydrology*, 353(3–4), 294–313.
<https://doi.org/10.1016/j.jhydrol.2008.02.015>

- Cobb, P. M. (1981). *The distribution and mechanisms of salt water intrusion in the fresh water aquifer and in Rattlesnake Creek, Stafford County, Kansas*. University of Kansas, Civil Engineering.
- Cogger, C. (1988). On-site Septic Systems: The risk of groundwater contamination. *Journal of Environmental Health*, 51(1), 12–16. JSTOR.
- Conn, K. E., Siegrist, R. L., Barber, L. B., & Meyer, M. T. (2009). Fate of trace organic compounds during vadose zone soil treatment in an onsite wastewater system. *Environmental Toxicology and Chemistry*, 29(2), 285–293. <https://doi.org/10.1002/etc.40>
- Copeland, C., Resources, & Division, I. (2006). *Animal waste and water quality: EPA regulation of concentrated animal feeding operations (CAFOs)*. <http://nationalaglawcenter.org/wp-content/uploads/assets/crs/RL31851.pdf>
- Craig, H. (1961). Isotopic Variations in Meteoric Waters. *Science*, 133(3465), 1702–1703. <https://doi.org/10.1126/science.133.3465.1702>
- Craun, G. F., Brunkard, J. M., Yoder, J. S., Roberts, V. A., Carpenter, J., Wade, T., Calderon, R. L., Roberts, J. M., Beach, M. J., & Roy, S. L. (2010). Causes of Outbreaks Associated with Drinking Water in the United States from 1971 to 2006. *Clinical Microbiology Reviews*, 23(3), 507–528. <https://doi.org/10.1128/CMR.00077-09>
- Dao, P. U., Heuzard, A. G., Le, T. X. H., Zhao, J., Yin, R., Shang, C., & Fan, C. (2024). The impacts of climate change on groundwater quality: A review. *Science of The Total Environment*, 912, 169241. <https://doi.org/10.1016/j.scitotenv.2023.169241>
- DeBaker, A., Kelley, G., Sax, W., Shafiee, F., Bussler, G., Hubbell, A., & Woodstock, H. (n.d.). *Cody Cook Rick Graham Steven Martin Jeff Postle Stanley Senger*.

- DeSimone, L. A. (2009). *Quality of water from domestic wells in principal aquifers of the United States, 1991-2004*. US Geological Survey. <https://pubs.usgs.gov/publication/sir20085227>
- Dubrovsky, N. M., Burow, K. R., Clark, G. M., Gronberg, J. M., Hamilton, P. A., Hitt, K. J., Mueller, D. K., Munn, M. D., Nolan, B. T., & Puckett, L. J. (2010). *The quality of our nation's waters: Nutrients in the nation's streams and groundwater, 1992-2004*. US Geological Survey.
- Euclidean Distance (Spatial Analyst)—ArcMap | Documentation*. (2021).
<https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-analyst-toolbox/euclidean-distance.htm>
- Fader, S. W., & Stullken, L. E. (1978). *Geohydrology of the Great Bend Prairie, Southcentral Kansas* (Vol. 4). University of Kansas Publications.
https://books.google.com/books?hl=en&lr=&id=N1BggVJRKpUC&oi=fnd&pg=PA1&dq=South+-Central+Kansas++Stuart+W.+Fader++Lloyd+E.+Stullken&ots=Qxl_YVuMn&sig=yK1dYS349TNwc1PIFf0R-sl7DMU
- feng, C., WANG, H., LU, N., CHEN, T., HE, H., LU, Y., & TU, X. M. (2014). Log-transformation and its implications for data analysis. *Shanghai Archives of Psychiatry*, 26(2), 105–109.
<https://doi.org/10.3969/j.issn.1002-0829.2014.02.009>
- Fewtrell, L. (2004). Drinking-Water Nitrate, Methemoglobinemia, and Global Burden of Disease: A Discussion. *Environmental Health Perspectives*, 112(14), 1371–1374.
<https://doi.org/10.1289/ehp.7216>
- Flanagan, S. V., Marvinney, R. G., & Zheng, Y. (2015). Influences on domestic well water testing behavior in a Central Maine area with frequent groundwater arsenic occurrence. *Science of The Total Environment*, 505, 1274–1281. <https://doi.org/10.1016/j.scitotenv.2014.05.017>

- Fontenot, B. E., Hunt, L. R., Hildenbrand, Z. L., Carlton Jr., D. D., Oka, H., Walton, J. L., Hopkins, D., Osorio, A., Bjorndal, B., Hu, Q. H., & Schug, K. A. (2013). An Evaluation of Water Quality in Private Drinking Water Wells Near Natural Gas Extraction Sites in the Barnett Shale Formation. *Environmental Science & Technology*, 47(17), 10032–10040.
<https://doi.org/10.1021/es4011724>
- Freeman, J. T. (2007). The use of bromide and chloride mass ratios to differentiate salt-dissolution and formation brines in shallow groundwaters of the Western Canadian Sedimentary Basin. *Hydrogeology Journal*, 15(7), 1377–1385. <https://doi.org/10.1007/s10040-007-0201-1>
- Gannon, R. S., Landon, M. K., Kulongoski, J. T., Stephens, M. J., Ball, L. B., Warden, J. G., Davis, T. A., Gillespie, J. M., & Cozzarelli, I. M. (2025). Relations of groundwater quality to long-term surface disposal of produced water near the Midway-Sunset and Buena Vista Oil Fields, California, USA. *Science of The Total Environment*, 987, 179637.
<https://doi.org/10.1016/j.scitotenv.2025.179637>
- Garcia, V., Cooter, E., Crooks, J., Hinckley, B., Murphy, M., & Xing, X. (2017). Examining the impacts of increased corn production on groundwater quality using a coupled modeling system. *Science of The Total Environment*, 586, 16–24. <https://doi.org/10.1016/j.scitotenv.2017.02.009>
- Geilfus, C.-M. (2019). Chloride in soil: From nutrient to soil pollutant. *Environmental and Experimental Botany*, 157, 299–309. <https://doi.org/10.1016/j.envexpbot.2018.10.035>
- Goldberg, V. M. (1989). Groundwater pollution by nitrates from livestock wastes. *Environmental Health Perspectives*, 83, 25.
- Guo, J., Wang, F., Vogt, R. D., Zhang, Y., & Liu, C.-Q. (2015). Anthropogenically enhanced chemical weathering and carbon evasion in the Yangtze Basin. *Scientific Reports*, 5(1), 11941.
<https://doi.org/10.1038/srep11941>

- Hamilton, P. A., & Helsel, D. R. (1995). Effects of Agriculture on Ground-Water Quality in Five Regions of the United States. *Groundwater*, 33(2), 217–226. <https://doi.org/10.1111/j.1745-6584.1995.tb00276.x>
- Hamlin, Q. F., Martin, S. L., Kendall, A. D., & Hyndman, D. W. (2022). Examining Relationships Between Groundwater Nitrate Concentrations in Drinking Water and Landscape Characteristics to Understand Health Risks. *GeoHealth*, 6(5), e2021GH000524. <https://doi.org/10.1029/2021GH000524>
- Hao, W., Liu, H., Hao, S., & Mao, K. (2024). Characterization of heavy metal contamination in groundwater of typical mining area in Hunan Province. *Scientific Reports*, 14(1), 13054. <https://doi.org/10.1038/s41598-024-63460-7>
- Harter, T., Kourakos, G., & Lockhart, K. (2014). Assessing potential impacts of livestock management on groundwater. *Nicholas Institute for Environmental Policy Solutions NI*, 14(03). http://nicholasinstitute.duke.edu/sites/default/files/ni_r_14-03_sr2_final.pdf
- Hathaway, L. R., Galle, O. K., Waugh, T. C., & Dickey, H. P. (1979). Chemical quality of irrigation waters in Ford County and the Great Bend Prairie of Kansas. *Bulletin (Kansas Geological Survey)*, 1–41.
- He, S., Wu, J., Wang, D., & He, X. (2022). Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere*, 290, 133388. <https://doi.org/10.1016/j.chemosphere.2021.133388>
- Helena, B. (2000). Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Water Research*, 34(3), 807–816. [https://doi.org/10.1016/S0043-1354\(99\)00225-0](https://doi.org/10.1016/S0043-1354(99)00225-0)

- Hijmans, R. J., Brown, A., & Barbosa, M. (2020). *terra: Spatial Data Analysis* (p. 1.9-34) [Dataset].
<https://doi.org/10.32614/CRAN.package.terra>
- Hildenbrand, Z. L., Carlton, D. D., Fontenot, B. E., Meik, J. M., Walton, J. L., Thacker, J. B., Korlie, S., Shelor, C. P., Kadjo, A. F., Clark, A., Usenko, S., Hamilton, J. S., Mach, P. M., Verbeck, G. F., Hudak, P., & Schug, K. A. (2016). Temporal variation in groundwater quality in the Permian Basin of Texas, a region of increasing unconventional oil and gas development. *Science of The Total Environment*, 562, 906–913. <https://doi.org/10.1016/j.scitotenv.2016.04.144>
- Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., & Megown, K. (2015). Completion of the 2011 National Land Cover Database for the Conterminous United States – Representing a Decade of Land Cover Change Information. *Photogrammetric Engineering & Remote Sensing*, 81(5), 345–354.
[https://doi.org/10.1016/S0099-1112\(15\)30100-2](https://doi.org/10.1016/S0099-1112(15)30100-2)
- Hussain, M. Z., Bhardwaj, A. K., Basso, B., Robertson, G. P., & Hamilton, S. K. (2019). Nitrate Leaching from Continuous Corn, Perennial Grasses, and Poplar in the US Midwest. *Journal of Environmental Quality*, 48(6), 1849–1855. <https://doi.org/10.2134/jeq2019.04.0156>
- John, B. W., & Gronberg, J. A. M. (2017). *County-Level Estimates of Nitrogen and Phosphorus from Commercial Fertilizer for the Conterminous United States, 1987-2012* [Dataset]. U.S. Geological Survey. <https://doi.org/10.5066/F7H41PKX>
- Johnson, G. R. (2022). PFAS in soil and groundwater following historical land application of biosolids. *Water Research*, 211, 118035. <https://doi.org/10.1016/j.watres.2021.118035>
- Johnson, S. L., Adams, R. M., & Perry, G. M. (1989). *Strategies for Reducing Groundwater Pollution for Agriculture: The Case of Irrigated Production in Oregon*.
<https://ageconsearch.umn.edu/record/270528/files/aaea-1989-064.pdf>

- Jolliffe, I. T., & Cadima, J. (2016a). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Jolliffe, I. T., & Cadima, J. (2016b). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Jones, A. Q., Dewey, C. E., Doré, K., Majowicz, S. E., McEwen, S. A., David, W.-T., Eric, M., Carr, D. J., & Henson, S. J. (2006). Public perceptions of drinking water: A postal survey of residents with private water supplies. *BMC Public Health*, 6(1), 94. <https://doi.org/10.1186/1471-2458-6-94>
- Juan Carlos González García. (2025). Impact of Agricultural Pollution on Groundwater Quality: A Hydrogeochemical Approach. *Journal of Information Systems Engineering and Management*, 10(43s), 768–775. <https://doi.org/10.52783/jisem.v10i43s.8478>
- Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5), 402–406. <https://doi.org/10.4097/kjae.2013.64.5.402>
- KDHE. (2017, November). *Livestock Waste Management Program*. Kansas Department of Health and Environment.
- KGS--Irrigation Series 4—Geohydrology of the Great Bend Prairie, South-Central Kansas. (1978). <https://www.kgs.ku.edu/Publications/Bulletins/IRR4/>
- Kheirfam, H., & Vafakhah, M. (2015). Assessment of some homogeneous methods for the regional analysis of suspended sediment yield in the south and southeast of the Caspian Sea. *Journal of Earth System Science*, 124(6), 1247–1263. <https://doi.org/10.1007/s12040-015-0604-7>

- Kim, H., & Kim, K. (2012). Microbial and chemical contamination of groundwater around livestock mortality burial sites in Korea—A review. *Geosciences Journal*, 16(4), 479–489.
<https://doi.org/10.1007/s12303-012-0036-1>
- Kirk, M. F. (with Open Textbook Library). (2024). *Microbiology for Earth Scientists*. New Prairie Press.
- Korsaeth, A., & Eltun, R. (2000). Nitrogen mass balances in conventional, integrated and ecological cropping systems and the relationship between balance calculations and nitrogen runoff in an 8-year field experiment in Norway. *Agriculture, Ecosystems & Environment*, 79(2–3), 199–214.
[https://doi.org/10.1016/S0167-8809\(00\)00129-8](https://doi.org/10.1016/S0167-8809(00)00129-8)
- Kreutzweiser, R., De Loë, R., Imgrund, K., Conboy, M. J., Simpson, H., & Plummer, R. (2011). Understanding stewardship behaviour: Factors facilitating and constraining private water well stewardship. *Journal of Environmental Management*, 92(4), 1104–1114.
<https://doi.org/10.1016/j.jenvman.2010.11.017>
- Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. <https://doi.org/10.1007/978-1-4614-6849-3>
- Kyte, E., Cey, E., Hrapovic, L., & Hao, X. (2023). Nitrate in shallow groundwater after more than four decades of manure application. *Journal of Contaminant Hydrology*, 256, 104200.
<https://doi.org/10.1016/j.jconhyd.2023.104200>
- Lane, A. D., Kirk, M. F., Whittemore, D. O., Stotler, R., Hildebrand, J., & Feril, O. (2020). Long-term (1970s–2016) changes in groundwater geochemistry in the High Plains aquifer in south-central Kansas, USA. *Hydrogeology Journal*, 28(2), 491–501. <https://doi.org/10.1007/s10040-019-02083-z>

- Lasagna, M., De Luca, D. A., Debernardi, L., & Clemente, P. (2013). Effect of the dilution process on the attenuation of contaminants in aquifers. *Environmental Earth Sciences*, 70(6), 2767–2784.
<https://doi.org/10.1007/s12665-013-2336-9>
- Layton, D. W., & Berry, D. W. (1973). Geology and ground-water resources of Pratt County, south-central Kansas. *Bulletin (Kansas Geological Survey)*, (205), 1–33.
- Lee, D., & Murphy, H. M. (2020). Private Wells and Rural Health: Groundwater Contaminants of Emerging Concern. *Current Environmental Health Reports*, 7(2), 129–139.
<https://doi.org/10.1007/s40572-020-00267-4>
- Lewis Jr, E. (2025). *BIG BEND GROUNDWATER MANAGEMENT DISTRICT NUMBER FIVE*.
- Li, P., Karunanidhi, D., Subramani, T., & Srinivasamoorthy, K. (2021). Sources and Consequences of Groundwater Contamination. *Archives of Environmental Contamination and Toxicology*, 80(1), 1–10. <https://doi.org/10.1007/s00244-020-00805-z>
- Li, Q., Zhang, H., Guo, S., Fu, K., Liao, L., Xu, Y., & Cheng, S. (2020). Groundwater pollution source apportionment using principal component analysis in a multiple land-use area in southwestern China. *Environmental Science and Pollution Research*, 27(9), 9000–9011.
<https://doi.org/10.1007/s11356-019-06126-6>
- Liu, C.-W., Lin, C.-N., Jang, C.-S., Ling, M.-P., & Tsai, J.-W. (2011). Assessing nitrate contamination and its potential health risk to Kinmen residents. *Environmental Geochemistry and Health*, 33(5), 503–514. <https://doi.org/10.1007/s10653-010-9367-x>
- Lockhart, K. M., King, A. M., & Harter, T. (2013). Identifying sources of groundwater nitrate contamination in a large alluvial groundwater basin with highly diversified intensive agricultural production. *Journal of Contaminant Hydrology*, 151, 140–154.
<https://doi.org/10.1016/j.jconhyd.2013.05.008>

- Maadooliat, M., Huang, J. Z., & Hu, J. (2015). Integrating Data Transformation in Principal Components Analysis. *Journal of Computational and Graphical Statistics*, 24(1), 84–103.
<https://doi.org/10.1080/10618600.2014.891461>
- Macfarlane, P. A., & Ackerman, J. (1983). *SOUNDINGS IN THE GREAT BEND PRAIRIE, SOUTH-CENTRAL KANSAS*.
- Madhav, S., & Singh, P. (2021). *Groundwater Geochemistry: Pollution and Remediation Methods*. John Wiley & Sons, Incorporated.
<http://ebookcentral.proquest.com/lib/ksu/detail.action?docID=6643625>
- Mbae, M., Hansen, P., Way, C., Mills, F., Willetts, J., Foster, T., & Evans, B. (2024). Onsite sanitation systems and contamination of groundwater: A systematic review of the evidence for risk using the source-pathway-receptor model. *PLOS Water*, 3(7), e0000167.
<https://doi.org/10.1371/journal.pwat.0000167>
- McAllister, T. A., & Topp, E. (2012). Role of livestock in microbiological contamination of water: Commonly the blame, but not always the source. *Animal Frontiers*, 2(2), 17–27.
- McLeod, L., Bharadwaj, L., Epp, T., & Waldner, C. (2017). Use of Principal Components Analysis and Kriging to Predict Groundwater-Sourced Rural Drinking Water Quality in Saskatchewan. *International Journal of Environmental Research and Public Health*, 14(9), 1065.
<https://doi.org/10.3390/ijerph14091065>
- Meena, V., Dotaniya, M. L., Saha, J. K., Das, H., & Patra, A. K. (2020). Impact of Lead Contamination on Agroecosystem and Human Health. In D. K. Gupta, S. Chatterjee, & C. Walther (Eds.), *Lead in Plants and the Environment* (pp. 67–82). Springer International Publishing. https://doi.org/10.1007/978-3-030-21638-2_4

- Menzie, C. A., Kashuba, R. O., & Goodfellow, W. L. (2020). Implications of sample size, rareness, and commonness for derivation of environmental benchmarks and criteria from field and laboratory data. *Ecotoxicology and Environmental Safety*, 190, 110117. <https://doi.org/10.1016/j.ecoenv.2019.110117>
- Mirzaee, S., & Mirzakhani Nafchi, A. (2025). Enhancing NUE in Corn Through Optimized Sensor-Based Prescription Maps. *Sensors*, 25(10), 3148. <https://doi.org/10.3390/s25103148>
- Murgulet, D., & Tick, G. R. (2013). Understanding the sources and fate of nitrate in a highly developed aquifer system. *Journal of Contaminant Hydrology*, 155, 69–81. <https://doi.org/10.1016/j.jconhyd.2013.09.004>
- N. K. Patni & J. L. B. Culley. (1989). Corn Silage Yield, Shallow Groundwater Quality and Soil Properties Under Different Methods and Times of Manure Application. *Transactions of the ASAE*, 32(6), 2123. <https://doi.org/10.13031/2013.31273>
- Naily, W. & Sudaryanto. (2018). Cl/Br Ratio to Determine Groundwater Quality. *IOP Conference Series: Earth and Environmental Science*, 118, 012020. <https://doi.org/10.1088/1755-1315/118/1/012020>
- Nguegang, B., & Ambushe, A. A. (2024a). Insight into the chemical and biochemical mechanisms governing inorganic contaminants removal by selective precipitation and neutralization in acid mine drainage treatment using MgO: A comparative study. *Journal of Water Process Engineering*, 59, 104924. <https://doi.org/10.1016/j.jwpe.2024.104924>
- Nguegang, B., & Ambushe, A. A. (2024b). Sustainable acid mine drainage treatment: A comprehensive review of passive, combined, and emerging technologies. *Environmental Engineering Research*, 30(4), 240592–0. <https://doi.org/10.4491/eer.2024.592>

- Nolan, B. T., Ruddy, B. C., Hitt, K. J., & Helsel, D. R. (1997). Risk of Nitrate in Groundwaters of the United States A National Perspective. *Environmental Science & Technology*, 31(8), 2229–2236.
<https://doi.org/10.1021/es960818d>
- On Bartlett’s sphericity test for correlation. (2022, April 27). *The DO Loop*.
<https://blogs.sas.com/content/iml/2022/04/27/bartletts-sphericity-test.html>
- Oviedo-Vargas, D., Anton, J., Coleman-Kammula, S., & Qin, X. (2025). Quantification of PFAS in soils treated with biosolids in ten northeastern US farms. *Scientific Reports*, 15(1), 5582.
<https://doi.org/10.1038/s41598-025-90184-z>
- Palandri, C. (2025). *Water Contamination from Industrial Livestock Production*.
https://clairepalandri.github.io/docs/wps/wp_afos-wq_CPalandri.pdf
- Perrin, A.-S., Probst, A., & Probst, J.-L. (2008). Impact of nitrogenous fertilizers on carbonate dissolution in small agricultural catchments: Implications for weathering CO₂ uptake at regional and global scales. *Geochimica et Cosmochimica Acta*, 72(13), 3105–3123.
<https://doi.org/10.1016/j.gca.2008.04.011>
- Preethi, B., Subramani, T., Karunanidhi, D., & Berhe, B. A. (2025). Evaluation of sulphate contamination in surface and groundwater with health hazards in the largest opencast lignite mine region of Asia. *Scientific Reports*, 16(1), 2962. <https://doi.org/10.1038/s41598-025-32763-8>
- Private well water data | Colorado Environmental Public Health Tracking*. (2026).
<https://coepht.colorado.gov/private-well-water-data>
- Probst, P., & Boulesteix, A.-L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, 18(181), 1–18.

- Probst, P., Wright, M. N., & Boulesteix, A.-L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3), e1301.
<https://doi.org/10.1002/widm.1301>
- Proger, S., Ginsburg, E., & Magendantz, H. (1942). The effects of the ingestion of excessive amounts of sodium chloride and water on patients with heart disease. *American Heart Journal*, 23(4), 555–566. [https://doi.org/10.1016/S0002-8703\(42\)90299-0](https://doi.org/10.1016/S0002-8703(42)90299-0)
- R: Common data transformations and standardizations. (2025, June 25). <https://search.r-project.org/CRAN/refmans/analogue/html/tran.html>
- Rajurkar, P. K. (2025). Groundwater Contamination From Industrial Waste And Its Implications For Sustainable Water Resource Management: Challenges And Emerging Remediation Pathways. *International Journal of Environmental Sciences*, 1250–1258. <https://doi.org/10.64252/512hg875>
- Ranganathan, P., & Hunsberger, S. (2024). Handling missing data in research. *Perspectives in Clinical Research*, 15(2), 99–101. https://doi.org/10.4103/picr.picr_38_24
- Rao, N. S. (2014). Spatial control of groundwater contamination, using principal component analysis. *Journal of Earth System Science*, 123(4), 715–728. <https://doi.org/10.1007/s12040-014-0430-3>
- Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, 47(1), 31–39.
<https://doi.org/10.17849/in-sm-47-01-31-39.1>
- Rivett, M. O., Buss, S. R., Morgan, P., Smith, J. W. N., & Bemment, C. D. (2008). Nitrate attenuation in groundwater: A review of biogeochemical controlling processes. *Water Research*, 42(16), 4215–4232. <https://doi.org/10.1016/j.watres.2008.07.020>
- Ryker, S. J., & Jones, J. L. (1995, October). *USGS OFR-95-445: Nitrate Concentrations in Ground Water of the Central Columbia Plateau*. <https://wa.water.usgs.gov/pubs/ofr/ofr95-445/>

- Sahoo, P. K., Kim, K., & Powell, M. A. (2016). Managing Groundwater Nitrate Contamination from Livestock Farms: Implication for Nitrate Management Guidelines. *Current Pollution Reports*, 2(3), 178–187. <https://doi.org/10.1007/s40726-016-0033-5>
- Sankhla, M. S. (2019). Contaminant of Heavy Metals in Groundwater & its Toxic Effects on Human Health & Environment. *International Journal of Environmental Sciences & Natural Resources*, 18(5). <https://doi.org/10.19080/IJESNR.2019.18.555996>
- Sarkar, S., Mukherjee, A., Gupta, S. D., Bhanja, S. N., & Bhattacharya, A. (2022). Predicting regional-scale elevated groundwater nitrate contamination risk using machine learning on natural and human-induced factors. *Acs Es&T Engineering*, 2(4), 689–702.
- Schader, L. M., Song, W., Kempker, R., & Benkeser, D. (2024). Don't let your analysis go to seed: On the impact of random seed on machine learning-based causal inference. *Epidemiology (Cambridge, Mass.)*, 35(6), 764–778. <https://doi.org/10.1097/EDE.0000000000001782>
- Schaider, L. A., Ackerman, J. M., & Rudel, R. A. (2016). Septic systems as sources of organic wastewater compounds in domestic drinking water wells in a shallow sand and gravel aquifer. *Science of The Total Environment*, 547, 470–481. <https://doi.org/10.1016/j.scitotenv.2015.12.081>
- Schaider, L. A., Rudel, R. A., Ackerman, J. M., Dunagan, S. C., & Brody, J. G. (2014). Pharmaceuticals, perfluorosurfactants, and other organic wastewater compounds in public drinking water wells in a shallow sand and gravel aquifer. *Science of The Total Environment*, 468–469, 384–393. <https://doi.org/10.1016/j.scitotenv.2013.08.067>
- Shapiro, A. M., & Falcone, J. A. (2022). Mapping areas of groundwater susceptible to transient contamination events from rapid infiltration into shallow fractured-rock aquifers in agricultural regions of the conterminous United States. In *Open-File Report* (Nos. 2022–1093). U.S. Geological Survey. <https://doi.org/10.3133/ofr20221093>

- Shrestha, A., & Luo, W. (2017). Analysis of Groundwater Nitrate Contamination in the Central Valley: Comparison of the Geodetector Method, Principal Component Analysis and Geographically Weighted Regression. *ISPRS International Journal of Geo-Information*, 6(10), 297. <https://doi.org/10.3390/ijgi6100297>
- Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling & Software*, 22(4), 464–475. <https://doi.org/10.1016/j.envsoft.2006.02.001>
- Singhal, B. B. S., & Gupta, R. P. (2010). Groundwater Contamination. In B. B. S. Singhal & R. P. Gupta, *Applied Hydrogeology of Fractured Rocks* (pp. 221–236). Springer Netherlands. https://doi.org/10.1007/978-90-481-8799-7_12
- Sinha, D., & Prasad, P. (2020). Health effects inflicted by chronic low-level arsenic contamination in groundwater: A global public health challenge. *Journal of Applied Toxicology*, 40(1), 87–131. <https://doi.org/10.1002/jat.3823>
- Sivakumar, M., Parthasarathy, S., & Padmapriya, T. (2024). Trade-off between training and testing ratio in machine learning for medical image processing. *PeerJ Computer Science*, 10, e2245. <https://doi.org/10.7717/peerj-cs.2245>
- Smith, A. H., & Steinmaus, C. M. (2011). Arsenic in drinking water. *BMJ*, 342(may05 2), d2248–d2248. <https://doi.org/10.1136/bmj.d2248>
- Sophocleous, M., & Ma, T. (1998). A decision support model to assess vulnerability to salt water intrusion in the great bend prairie aquifer of Kansas. *Groundwater*, 36(3), 476–483.
- Spalding, R. F., & Exner, M. E. (1993). Occurrence of Nitrate in Groundwater—A Review. *Journal of Environmental Quality*, 22(3), 392–402. <https://doi.org/10.2134/jeq1993.00472425002200030002x>

- Speiser, J. L., Durkalski, V. L., & Lee, W. M. (2015). Random forest classification of etiologies for an orphan disease. *Statistics in Medicine*, 34(5), 887–899. <https://doi.org/10.1002/sim.6351>
- Speiser, J. L., Miller, M. E., Tooze, J., & Ip, E. (2019). A comparison of random forest variable selection methods for classification prediction modeling. *Expert Systems with Applications*, 134, 93–101. <https://doi.org/10.1016/j.eswa.2019.05.028>
- Stillo, F., & MacDonald Gibson, J. (2017). Exposure to Contaminated Drinking Water and Health Disparities in North Carolina. *American Journal of Public Health*, 107(1), 180–185. <https://doi.org/10.2105/AJPH.2016.303482>
- Sun, H., Zheng, W., Wang, S., Ma, L., Min, L., & Shen, Y. (2024). Variation of nitrate sources affected by precipitation with different intensities in groundwater in the piedmont plain area of alluvial-pluvial fan. *Journal of Environmental Management*, 367, 121885. <https://doi.org/10.1016/j.jenvman.2024.121885>
- Suttles, K. M., Smoliak, B. V., Ranade, A. P., Potter, S. F., Jaeger, M., & McLellan, E. L. (2024). Kansas agriculture in 2050: A pathway for climate-resilient crop production. *Frontiers in Sustainable Food Systems*, 8, 1404315. <https://doi.org/10.3389/fsufs.2024.1404315>
- Swartz, C. H., Reddy, S., Benotti, M. J., Yin, H., Barber, L. B., Brownawell, B. J., & Rudel, R. A. (2006). Steroid Estrogens, Nonylphenol Ethoxylate Metabolites, and Other Wastewater Contaminants in Groundwater Affected by a Residential Septic System on Cape Cod, MA. *Environmental Science & Technology*, 40(16), 4894–4902. <https://doi.org/10.1021/es052595+>
- Swistock, B. R., Clemens, S., Sharpe, W. E., & Rummel, S. (2013). Water Quality and Management of Private Drinking Water Wells in Pennsylvania. *Journal of Environmental Health*, 75(6), 60–67. JSTOR.

- Tillman, F. D., Beisner, K. R., & Jones, C. J. R. (2023). Arsenic in groundwater in the Grand Canyon region and an evaluation of potential pathways for arsenic contamination of groundwater from breccia pipe uranium mining. *PLOS Water*, 2(6), e0000109.
<https://doi.org/10.1371/journal.pwat.0000109>
- Townsend, M. A., Sleezer, R. O., & Macko, S. A. (1996). Effects of agricultural practices and vadose zone stratigraphy on nitrate concentration in ground water in Kansas, USA. *Water Science and Technology*, 33(4–5), 219–226. ProQuest One Academic (1943330596).
- Townsend, M. A., & Young, D. P. (1995). Factors Affecting Nitrate Concentrations in Ground Water in Stafford County, Kansas. *Current Research in Earth Sciences*, 1–9.
<https://doi.org/10.17161/cres.v0i238.11844>
- Townsend, M. A., & Young, D. P. (2000). Assessment of Nitrate—Nitrogen Distribution in Kansas Groundwater, 1990–1998. *Natural Resources Research*, 9(2), 125–134.
<https://doi.org/10.1023/A:1010143426576>
- Tsang, J., Everard, M., & Staddon, C. (2026). “Forever Chemicals in Our Water!”: Prevalence of PFAS in Drinking Water, Associated Health Risks, and Approaches to Regulation. *WIREs Water*, 13(1), e70057. <https://doi.org/10.1002/wat2.70057>
- U.S. Geological Survey. (2023). *County-Level Estimates of Nitrogen and Phosphorus from Commercial Fertilizer for the Conterminous United States, 1987-2006* [Dataset]. U.S. Geological Survey. <https://doi.org/10.5066/P9E60J5U>
- USDA - National Agricultural Statistics Service—Surveys—Agricultural Chemical Use Program. (2026). https://www.nass.usda.gov/Surveys/Guide_to_NASS_Surveys/Chemical_Use/
- USDA, N. (2021). USDA national agricultural statistics service cropland data layer. *Publ. Crop. Data Layer*. Retrieved, 1.

- Vafakhah, M., & Janizadeh, S. (2021). Application of artificial neural network and adaptive neuro-fuzzy inference system in streamflow forecasting. In *Advances in Streamflow Forecasting* (pp. 171–191). Elsevier. <https://doi.org/10.1016/B978-0-12-820673-7.00002-0>
- Wagner, C. R., Fitzgerald, S. A., McSwain, K. B., Harden, S. L., Gurley, L. N., & Rogers, S. W. (2015). Effect of land-applied biosolids on surface-water nutrient yields and groundwater quality in Orange County, North Carolina. In *Scientific Investigations Report* (Nos. 2014–5240). U.S. Geological Survey. <https://doi.org/10.3133/sir20145240>
- Wallender, E. K., Ailes, E. C., Yoder, J. S., Roberts, V. A., & Brunkard, J. M. (2014). Contributing Factors to Disease Outbreaks Associated with Untreated Groundwater. *Groundwater*, 52(6), 886–897. <https://doi.org/10.1111/gwat.12121>
- Wells, N. S., Clough, T. J., Johnson-Beebout, S. E., Elberling, B., & Baisden, W. T. (2019). Effects of denitrification and transport on the isotopic composition of nitrate ($\delta^{18}\text{O}$, $\delta^{15}\text{N}$) in freshwater systems. *Science of The Total Environment*, 651, 2228–2234. <https://doi.org/10.1016/j.scitotenv.2018.10.065>
- Wheeler, D. C., Nolan, B. T., Flory, A. R., DellaValle, C. T., & Ward, M. H. (2015). Modeling groundwater nitrate concentrations in private wells in Iowa. *Science of The Total Environment*, 536, 481–488. <https://doi.org/10.1016/j.scitotenv.2015.07.080>
- Whittemore, D. O., Butler Jr, J. J., & Wilson, B. B. (2018). Status of the High Plains aquifer in Kansas. *Kansas Geological Survey Technical Series*, 22, 1–28.
- Wick, K., Heumesser, C., & Schmid, E. (2012). Groundwater nitrate contamination: Factors and indicators. *Journal of Environmental Management*, 111, 178–186. <https://doi.org/10.1016/j.jenvman.2012.06.030>

Winter, T. C., Harvey, J. W., Franke, O. L., & Alley, W. M. (1998). Ground water and surface water: A single resource. In *Circular* (No. 1139). U.S. Geological Survey.

<https://doi.org/10.3133/cir1139>

Wisconsin Groundwater Coordinating Council Wisconsin. Department of Natural Resources. (2021).

Wisconsin Groundwater Coordinating Council: Report to the legislature fiscal year 2021.

<https://www.wistatedocuments.org/digital/collection/p267601coll4/id/30293/>

World Health Organization. (2009). *Calcium and magnesium in drinking-water: Public health significance*. WHO.

<https://books.google.com/books?hl=en&lr=&id=dInKO6fSQqMC&oi=fnd&pg=PP2&dq=effects+of+excess+calcium+and+magnesium+in+drinking+water%5C&ots=IQlIMNWUod&sig=-VEoyTjalawvKM0-DROleFJvXCk>

Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77, 1–17.

<https://doi.org/10.18637/jss.v077.i01>

Xu, Y., Jakobsson, K., Harari, F., Andersson, E. M., & Li, Y. (2023). Exposure to high levels of PFAS through drinking water is associated with increased risk of type 2 diabetes—Findings from a register-based study in Ronneby, Sweden. *Environmental Research*, 225, 115525.

<https://doi.org/10.1016/j.envres.2023.115525>

Xue, D., Botte, J., De Baets, B., Accoe, F., Nestler, A., Taylor, P., Van Cleemput, O., Berglund, M., & Boeckx, P. (2009). Present limitations and future prospects of stable isotope methods for nitrate source identification in surface- and groundwater. *Water Research*, 43(5), 1159–1170.

<https://doi.org/10.1016/j.watres.2008.12.048>

- Zhou, S., Xiong, H., Song, X., He, Z., Peng, Q., Guo, X., Xiong, R., Yang, S., Ma, C., & Wang, Y. (2025). Spatial prediction of groundwater nitrate by explainable machine learning models and remote sensing indicators under conditions of imbalanced datasets. *Geocarto International*, 40(1), 2557953. <https://doi.org/10.1080/10106049.2025.2557953>
- Zirkle, K. W., Nolan, B. T., Jones, R. R., Weyer, P. J., Ward, M. H., & Wheeler, D. C. (2016). Assessing the relationship between groundwater nitrate and animal feeding operations in Iowa (USA). *Science of The Total Environment*, 566–567, 1062–1068. <https://doi.org/10.1016/j.scitotenv.2016.05.130>

Appendix A – Random Forest Model

```
library(tidyverse)
library(readxl)
library(janitor)
library(caret)
library(ranger)
library(pROC)

input_file <- "D:/StatAnalysis/RFTS2/step8.xlsx"

out_dir <- "D:/StatAnalysis/RFTS2/RF_classification4"

if (!dir.exists(out_dir)) {

  dir.create(
    out_dir,
    recursive = TRUE
  )
}

cleaned_names <- read_excel(input_file) %>%
  clean_names() %>%
  names()

print(cleaned_names)

# Predictor variables

predictors <- c(
  "som",
  "o2_mg_l",
  "p_h",
  "cultivated_10",
  "herbaceous_10",
  "forest_10",
  "herbaceous_wetlands_10",
  "satthick_2023_2025",
  "predev_dtw_updated",
  "dtw_2023_2025",
  "ksat",
  "avg_pprec",
  "corn_10",
  "barren_10",
  "sum_nfert",
  "urban_m",
```

```

    "kdau_1",
    "all_streams_m"
  )

# Read and prepare the data

model_data <- read_excel(input_file) %>%
  clean_names() %>%
  select(
    nitrate,
    all_of(predictors)
  ) %>%
  mutate(
    across(
      everything(),
      as.numeric
    )
  ) %>%
  filter(
    !is.na(nitrate),
    is.finite(nitrate)
  ) %>%
  mutate(
    nitrate_risk = factor(
      ifelse(
        nitrate >= 10,
        "High",
        "Low"
      ),
      levels = c(
        "Low",
        "High"
      )
    )
  )

cat("\nTotal number of wells:", nrow(model_data), "\n\n")

cat("Number of wells in each nitrate-risk group:\n")

print(
  table(model_data$nitrate_risk)
)
# Random Forest formula
rf_formula <- as.formula(
  paste(

```

```

    "nitrate_risk ~",
    paste(
      predictors,
      collapse = " + "
    )
  )
)

# hyperparameter tuning settings
tuning_grid <- expand.grid(

  mtry = c(
    3,
    4,
    5,
    6,
    8
  ),

  min_node_size = c(
    5,
    6,
    7,
    8,
    9,
    10,
    11,
    12,
    13,
    14,
    15,
    16,
    17,
    18,
    19,
    20,
    21,
    22,
    23,
    24,
    25
  )
)

# Candidate forest sizes for the automated tree-convergence analysis

```

```

tree_values <- c(
  800,
  900,
  1000,
  1100,
  1200,
  1300,
  1400,
  1500
)

cat("\nHyperparameter tuning grid:\n")

print(tuning_grid)

cat(
  "\nNumber of combinations:",
  nrow(tuning_grid),
  "\n"
)
# replace missing values
impute_training_and_test <- function(
  training_data,
  testing_data,
  predictor_names
) {

  training_medians <- sapply(
    training_data[predictor_names],
    function(x) {

      median(
        x,
        na.rm = TRUE
      )
    }
  )

  if (any(!is.finite(training_medians))) {

    stop(
      "At least one predictor contains no usable training values."
    )
  }

  for (variable in predictor_names) {

```

```

training_missing <- is.na(
  training_data[[variable]]
) |
!is.finite(
  training_data[[variable]]
)

testing_missing <- is.na(
  testing_data[[variable]]
) |
!is.finite(
  testing_data[[variable]]
)

training_data[[variable]][training_missing] <-
  training_medians[[variable]]

testing_data[[variable]][testing_missing] <-
  training_medians[[variable]]
}

return(
  list(
    training_data = training_data,
    testing_data = testing_data,
    training_medians = training_medians
  )
)
}

# automated tree-convergence analysis
set.seed(321)

tree_train_index <- createDataPartition(
  model_data$nitrate_risk,
  p = 0.70,
  list = FALSE
)

tree_train_data <- model_data[
  tree_train_index,
]

tree_test_data <- model_data[
  -tree_train_index,
]

```

```

]

tree_imputation <- impute_training_and_test(
  training_data = tree_train_data,
  testing_data = tree_test_data,
  predictor_names = predictors
)

tree_train_data <- tree_imputation$training_data

tree_mtry <- floor(
  sqrt(
    length(predictors)
  )
)

tree_min_node_size <- 10

tree_results <- vector(
  mode = "list",
  length = length(tree_values)
)

cat(
  "\nStarting automated tree-convergence analysis...\n\n"
)

for (i in seq_along(tree_values)) {

  current_number_of_trees <- tree_values[i]

  cat(
    "Testing",
    current_number_of_trees,
    "trees\n"
  )

  tree_model <- ranger(
    formula = rf_formula,
    data = tree_train_data,
    probability = TRUE,
    num.trees = current_number_of_trees,
    mtry = tree_mtry,
    min.node.size = tree_min_node_size,
    importance = "none",
    oob.error = TRUE,

```

```

    write.forest = FALSE,
    seed = 321
  )

  tree_oob_probability <-
    tree_model$predictions[, "High"]

  valid_tree_rows <-
    !is.na(tree_oob_probability) &
    is.finite(tree_oob_probability)

  tree_roc <- roc(
    response =
      tree_train_data$nitrate_risk[valid_tree_rows],
    predictor =
      tree_oob_probability[valid_tree_rows],
    levels = c(
      "Low",
      "High"
    ),
    direction = "<",
    quiet = TRUE
  )

  tree_results[[i]] <- tibble(
    Number_of_trees = current_number_of_trees,
    OOB_AUC = as.numeric(
      auc(tree_roc)
    )
  )
}

tree_convergence_results <- bind_rows(
  tree_results
)

maximum_tree_auc <- max(
  tree_convergence_results$OOB_AUC,
  na.rm = TRUE
)

tree_auc_tolerance <- 0.005

selected_tree_row <- tree_convergence_results %>%
  filter(
    OOB_AUC >=

```

```

    maximum_tree_auc -
    tree_auc_tolerance
  ) %>%
  arrange(
    Number_of_trees
  ) %>%
  slice(1)

selected_trees <- as.integer(
  selected_tree_row$Number_of_trees
)

# Use the selected number for both tuning and final models.

tuning_trees <- selected_trees
final_trees <- selected_trees

cat(
  "\nTree-convergence results:\n"
)

print(
  tree_convergence_results
)

cat(
  "\nMaximum OOB AUC:",
  round(maximum_tree_auc, 3),
  "\n"
)

cat(
  "Selected number of trees:",
  selected_trees,
  "\n\n"
)

write.csv(
  tree_convergence_results,
  file.path(
    out_dir,
    "tree_convergence_results.csv"
  ),
  row.names = FALSE
)

```

```

tree_convergence_plot <- ggplot(
  tree_convergence_results,
  aes(
    x = Number_of_trees,
    y = OOB_AUC
  )
) +
  geom_line(
    linewidth = 0.8
  ) +
  geom_point(
    size = 2
  ) +
  geom_vline(
    xintercept = selected_trees,
    linetype = "dashed"
  ) +
  labs(
    x = "Number of trees",
    y = "OOB AUC"
  ) +
  theme_classic(
    base_size = 12
  ) +
  theme(
    text = element_text(
      family = "Times New Roman"
    ),
    axis.text = element_text(
      color = "black"
    )
  )
)

print(
  tree_convergence_plot
)

ggsave(
  filename = file.path(
    out_dir,
    "tree_convergence_plot.png"
  ),
  plot = tree_convergence_plot,
  width = 7,
  height = 5,
  dpi = 600,

```

```

  bg = "white"
)

rm(
  tree_model,
  tree_train_data,
  tree_test_data,
  tree_imputation
)

gc()

# tuning random forest hyperparameters
tune_random_forest <- function(
  training_data,
  model_formula,
  parameter_grid,
  number_of_trees,
  base_seed
) {

  tuning_results <- vector(
    mode = "list",
    length = nrow(parameter_grid)
  )

  for (g in seq_len(nrow(parameter_grid))) {

    current_mtry <- parameter_grid$mtry[g]

    current_node_size <-
      parameter_grid$min_node_size[g]

    current_seed <- base_seed

    tuning_model <- ranger(
      formula = model_formula,
      data = training_data,
      probability = TRUE,
      num.trees = number_of_trees,
      mtry = current_mtry,
      min.node.size = current_node_size,
      importance = "none",
      oob.error = TRUE,
      write.forest = FALSE,
      seed = current_seed
    )
  }
}

```

```

)

oob_probability <-
  tuning_model$predictions[, "High"]

valid_rows <- !is.na(oob_probability) &
  is.finite(oob_probability)

tuning_roc <- roc(
  response =
    training_data$nitrate_risk[valid_rows],

  predictor =
    oob_probability[valid_rows],

  levels = c(
    "Low",
    "High"
  ),

  direction = "<",
  quiet = TRUE
)

tuning_results[[g]] <- tibble(
  mtry = current_mtry,
  min_node_size = current_node_size,
  OOB_AUC = as.numeric(
    auc(tuning_roc)
  )
)
}

tuning_results <- bind_rows(
  tuning_results
)

best_parameters <- tuning_results %>%
  arrange(
    desc(OOB_AUC),
    mtry,
    min_node_size
  ) %>%
  slice(1)

return(

```

```

    list(
      all_results = tuning_results,
      best_parameters = best_parameters
    )
  )
}

# repeated stratified train-test validation
n_repeats <- 100

repeated_results <- vector(
  mode = "list",
  length = n_repeats
)

# This list saves the tuning results from every repetition

all_tuning_results <- vector(
  mode = "list",
  length = n_repeats
)

cat(
  "\nStarting repeated train-test validation...\n\n"
)

for (i in seq_len(n_repeats)) {

  # Create a stratified 70/30 split

  split_seed <- 1000 + i

  set.seed(split_seed)

  train_index_i <- createDataPartition(
    model_data$nitrate_risk,
    p = 0.70,
    list = FALSE
  )

  train_i <- model_data[
    train_index_i,
  ]

  test_i <- model_data[
    -train_index_i,
  ]

```

```

]

# Impute missing values

imputed_i <- impute_training_and_test(
  training_data = train_i,
  testing_data = test_i,
  predictor_names = predictors
)

train_i <- imputed_i$training_data

test_i <- imputed_i$testing_data

#Tune hyperparameters using training data only
tuning_i <- tune_random_forest(
  training_data = train_i,
  model_formula = rf_formula,
  parameter_grid = tuning_grid,
  number_of_trees = tuning_trees,
  base_seed = 10000 + i * 100
)

tuning_results_i <- tuning_i$all_results

best_parameters_i <- tuning_i$best_parameters

best_mtry_i <- best_parameters_i$mtry

best_node_size_i <-
  best_parameters_i$min_node_size

best_tuning_auc_i <-
  best_parameters_i$OOB_AUC

# Save all tuning combinations for this repetition

all_tuning_results[[i]] <- tuning_results_i %>%
  mutate(
    Repetition = i,
    .before = 1
  )

# Fit the selected Random Forest
model_seed <- 5000 + i

```

```

rf_i <- ranger(
  formula = rf_formula,
  data = train_i,
  probability = TRUE,
  num.trees = final_trees,
  mtry = best_mtry_i,
  min.node.size = best_node_size_i,
  importance = "none",
  oob.error = TRUE,
  write.forest = TRUE,
  seed = model_seed
)

# oob evaluation for this repetition

train_i$oob_probability <-
  rf_i$predictions[, "High"]

oob_i <- train_i %>%
  filter(
    !is.na(oob_probability),
    is.finite(oob_probability)
  )

# OOB ROC and AUC

oob_roc_i <- roc(
  response = oob_i$nitrate_risk,
  predictor = oob_i$oob_probability,
  levels = c(
    "Low",
    "High"
  ),
  direction = "<",
  quiet = TRUE
)

oob_auc_i <- as.numeric(
  auc(oob_roc_i)
)

cutoff_i <- as.numeric(
  coords(
    oob_roc_i,
    x = "best",
    best.method = "youden",

```

```

    ret = "threshold",
    transpose = FALSE
  )
)

# independent test evaluation

test_probabilities_i <- predict(
  rf_i,
  data = test_i
)$predictions

test_i$probability_high <-
  test_probabilities_i[, "High"]

test_i$predicted_risk <- factor(
  ifelse(
    test_i$probability_high >= cutoff_i,
    "High",
    "Low"
  ),
  levels = c(
    "Low",
    "High"
  )
)

#Test ROC and AUC

test_roc_i <- roc(
  response = test_i$nitrate_risk,
  predictor = test_i$probability_high,
  levels = c(
    "Low",
    "High"
  ),
  direction = "<",
  quiet = TRUE
)

test_auc_i <- as.numeric(
  auc(test_roc_i)
)

#Test confusion matrix

```

```

test_cm_i <- confusionMatrix(
  data = test_i$predicted_risk,
  reference = test_i$nitrate_risk,
  positive = "High"
)

repeated_results[[i]] <- tibble(
  Repetition = i,
  Split_seed = split_seed,
  Model_seed = model_seed,
  Training_wells = nrow(train_i),
  Testing_wells = nrow(test_i),
  Selected_mtry = best_mtry_i,
  Selected_min_node_size = best_node_size_i,
  Tuning_OOB_AUC = best_tuning_auc_i,
  Final_OOB_AUC = oob_auc_i,
  OOB_cutoff = cutoff_i,
  Test_AUC = test_auc_i,

  Test_accuracy = unname(
    test_cm_i$overall["Accuracy"]
  ),

  Test_sensitivity = unname(
    test_cm_i$byClass["Sensitivity"]
  ),

  Test_specificity = unname(
    test_cm_i$byClass["Specificity"]
  ),

  Test_balanced_accuracy = unname(
    test_cm_i$byClass[
      "Balanced Accuracy"
    ]
  ),

  Test_precision = unname(
    test_cm_i$byClass[
      "Pos Pred Value"
    ]
  ),

  cat(

```

```

    "Completed repetition",
    i,
    "of",
    n_repeats,
    "\n"
  )
}

```

```

repeated_results_df <- bind_rows(
  repeated_results
)

```

```

all_tuning_results_df <- bind_rows(
  all_tuning_results
)

```

```

repeated_summary <- repeated_results_df %>%
  select(
    Tuning_OOB_AUC,
    Final_OOB_AUC,
    OOB_cutoff,
    Test_AUC,
    Test_accuracy,
    Test_sensitivity,
    Test_specificity,
    Test_balanced_accuracy,
    Test_precision,
  ) %>%
  pivot_longer(
    cols = everything(),
    names_to = "Metric",
    values_to = "Value"
  ) %>%
  group_by(Metric) %>%
  summarise(
    Mean = mean(
      Value,
      na.rm = TRUE
    ),
    SD = sd(
      Value,
      na.rm = TRUE
    ),

```

```

Minimum = min(
  Value,
  na.rm = TRUE
),

Median = median(
  Value,
  na.rm = TRUE
),

Maximum = max(
  Value,
  na.rm = TRUE
),

.groups = "drop"
) %>%
mutate(
  across(
    c(
      Mean,
      SD,
      Minimum,
      Median,
      Maximum
    ),
    ~ round(.x, 3)
  ),

Mean_SD = paste0(
  sprintf(
    "%.3f",
    Mean
  ),
  " ± ",
  sprintf(
    "%.3f",
    SD
  )
)
)

cat(
  "REPEATED VALIDATION SUMMARY\n"

```

```

)

print(repeated_summary)

parameter_selection_summary <-
  repeated_results_df %>%
  count(
    Selected_mtry,
    Selected_min_node_size,
    name = "Times_selected"
  ) %>%
  arrange(
    desc(Times_selected),
    Selected_mtry,
    Selected_min_node_size
  ) %>%
  mutate(
    Selection_percent = round(
      Times_selected /
      n_repeats *
      100,
      1
    )
  )

cat(
  "HYPERPARAMETER SELECTION FREQUENCY\n"
)

print(parameter_selection_summary)

write.csv(
  repeated_results_df,
  file.path(
    out_dir,
    "repeated_train_test_results.csv"
  ),
  row.names = FALSE
)

write.csv(
  repeated_summary,
  file.path(
    out_dir,
    "repeated_train_test_summary.csv"
  ),

```

```

    row.names = FALSE
  )

write.csv(
  parameter_selection_summary,
  file.path(
    out_dir,
    "hyperparameter_selection_summary.csv"
  ),
  row.names = FALSE
)

write.csv(
  all_tuning_results_df,
  file.path(
    out_dir,
    "all_hyperparameter_tuning_results.csv"
  ),
  row.names = FALSE
)

#Create the seed-123 train-test split

set.seed(123)

train_index <- createDataPartition(
  model_data$nitrate_risk,
  p = 0.70,
  list = FALSE
)

train_data <- model_data[
  train_index,
]

test_data <- model_data[
  -train_index,
]

final_imputation <- impute_training_and_test(
  training_data = train_data,
  testing_data = test_data,
  predictor_names = predictors
)

train_data <- final_imputation$training_data

```

```

test_data <- final_imputation$testing_data

# These medians can be used later for raster mapping

training_medians <-
  final_imputation$training_medians

seed123_tuning <- tune_random_forest(
  training_data = train_data,
  model_formula = rf_formula,
  parameter_grid = tuning_grid,
  number_of_trees = tuning_trees,
  base_seed = 20000
)

seed123_tuning_results <-
  seed123_tuning$all_results

seed123_best_parameters <-
  seed123_tuning$best_parameters

final_mtry <-
  seed123_best_parameters$mtry

final_min_node_size <-
  seed123_best_parameters$min_node_size

cat(
  "seed-123 selected parameters\n"
)

cat(
  "Selected number of trees:",
  selected_trees,
  "\n"
)

cat(
  "Selected mtry:",
  final_mtry,
  "\n"
)

cat(
  "Selected min.node.size:",

```

```

    final_min_node_size,
    "\n"
  )

cat(
  "Training OOB AUC during tuning:",
  round(
    seed123_best_parameters$OOB_AUC,
    3
  ),
  "\n"
)

# Save seed-123 tuning results

write.csv(
  seed123_tuning_results,
  file.path(
    out_dir,
    "seed123_hyperparameter_tuning.csv"
  ),
  row.names = FALSE
)

#Fit the final seed-123 Random Forest

rf_model <- ranger(
  formula = rf_formula,
  data = train_data,
  probability = TRUE,
  num.trees = final_trees,
  mtry = final_mtry,
  min.node.size = final_min_node_size,
  importance = "permutation",
  oob.error = TRUE,
  write.forest = TRUE,
  seed = 123
)

# FINAL-MODEL OOB EVALUATION

train_data$oob_probability <-
  rf_model$predictions[, "High"]

oob_data <- train_data %>%
  filter(

```

```

      !is.na(oob_probability),
      is.finite(oob_probability)
    )

oob_roc <- roc(
  response = oob_data$nitrate_risk,
  predictor = oob_data$oob_probability,
  levels = c(
    "Low",
    "High"
  ),
  direction = "<",
  quiet = TRUE
)

oob_auc <- as.numeric(
  auc(oob_roc)
)

optimal_cutoff <- as.numeric(
  coords(
    oob_roc,
    x = "best",
    best.method = "youden",
    ret = "threshold",
    transpose = FALSE
  )
)[1]

oob_data$oob_prediction <- factor(
  ifelse(
    oob_data$oob_probability >= optimal_cutoff,
    "High",
    "Low"
  ),
  levels = c(
    "Low",
    "High"
  )
)

oob_cm <- confusionMatrix(
  data = oob_data$oob_prediction,
  reference = oob_data$nitrate_risk,

```

```

    positive = "High"
  )

  oob_accuracy <- unname(
    oob_cm$overall["Accuracy"]
  )

  oob_sensitivity <- unname(
    oob_cm$byClass["Sensitivity"]
  )

  oob_specificity <- unname(
    oob_cm$byClass["Specificity"]
  )

  oob_balanced_accuracy <- unname(
    oob_cm$byClass["Balanced Accuracy"]
  )

  oob_precision <- unname(
    oob_cm$byClass["Pos Pred Value"]
  )

  # final-model independent test evaluation

  test_probabilities <- predict(
    rf_model,
    data = test_data
  )$predictions

  test_data$probability_high <-
    test_probabilities[, "High"]

  test_roc <- roc(
    response = test_data$nitrate_risk,
    predictor = test_data$probability_high,
    levels = c(
      "Low",
      "High"
    ),
    direction = "<",
    quiet = TRUE
  )

  test_auc <- as.numeric(
    auc(test_roc)
  )

```

```
)

test_data$predicted_risk <- factor(
  ifelse(
    test_data$probability_high >= optimal_cutoff,
    "High",
    "Low"
  ),
  levels = c(
    "Low",
    "High"
  )
)
```

```
#Test confusion matrix
```

```
test_cm <- confusionMatrix(
  data = test_data$predicted_risk,
  reference = test_data$nitrate_risk,
  positive = "High"
)
```

```
# Test metrics
```

```
test_accuracy <- unname(
  test_cm$overall["Accuracy"]
)
```

```
test_sensitivity <- unname(
  test_cm$byClass["Sensitivity"]
)
```

```
test_specificity <- unname(
  test_cm$byClass["Specificity"]
)
```

```
test_balanced_accuracy <- unname(
  test_cm$byClass["Balanced Accuracy"]
)
```

```
test_precision <- unname(
  test_cm$byClass["Pos Pred Value"]
)

)
```

```

# final-model training evaluation

train_probabilities <- predict(
  rf_model,
  data = train_data
)$predictions

train_data$probability_high <-
  train_probabilities[, "High"]

train_roc <- roc(
  response = train_data$nitrate_risk,
  predictor = train_data$probability_high,
  levels = c(
    "Low",
    "High"
  ),
  direction = "<",
  quiet = TRUE
)

train_auc <- as.numeric(
  auc(train_roc)
)

train_data$predicted_risk <- factor(
  ifelse(
    train_data$probability_high >= optimal_cutoff,
    "High",
    "Low"
  ),
  levels = c(
    "Low",
    "High"
  )
)

# Training confusion matrix

train_cm <- confusionMatrix(
  data = train_data$predicted_risk,
  reference = train_data$nitrate_risk,
  positive = "High"
)

```

```

# Training metrics

train_accuracy <- unname(
  train_cm$overall["Accuracy"]
)

train_sensitivity <- unname(
  train_cm$byClass["Sensitivity"]
)

train_specificity <- unname(
  train_cm$byClass["Specificity"]
)

train_balanced_accuracy <- unname(
  train_cm$byClass["Balanced Accuracy"]
)

train_precision <- unname(
  train_cm$byClass["Pos Pred Value"]
)

train_f1 <- unname(
  train_cm$byClass["F1"]
)

train_kappa <- unname(
  train_cm$overall["Kappa"]
)

train_brier <- mean(
  (
    train_data$probability_high -
    ifelse(
      train_data$nitrate_risk == "High",
      1,
      0
    )
  )^2
)

# variable importance

variable_labels <- c(
  som = "Soil organic matter",
  o2_mg_l = "Dissolved oxygen",

```

```

p_h = "pH",
cultivated_10 = "Cultivated area",
herbaceous_10 = "Herbaceous area",
forest_10 = "Forest area",
herbaceous_wetlands_10 = "Herbaceous wetlands",
sathick_2023_2025 = "Saturated thickness",
predev_dtw_updated = "Predevelopment depth to water",
dtw_2023_2025 = "Current depth to water",
ksat = "Hydraulic conductivity",
avg_pprec = "Average precipitation",
corn_10 = "Corn area",
barren_10 = "Barren area",
sum_nfert = "Cumulative N fertilizer",
urban_m = "Distance to urban area",
kdau_1 = "Animal-unit density",
all_streams_m = "Distance to streams"
)

```

```

importance_df <- tibble(

  Variable = names(
    rf_model$variable.importance
  ),

  Importance = as.numeric(
    rf_model$variable.importance
  )

) %>%
mutate(
  Display_name = recode(
    Variable,
    !!!variable_labels
  )
) %>%
arrange(
  desc(Importance)
)

```

```

# final-model summary

```

```

summary_table <- tibble(

  Metric = c(
    "Selected number of trees",

```

```

"Selected mtry",
"Selected minimum node size",
"Training AUC",
"OOB AUC",
"Test AUC",
"OOB-derived cutoff",
"Training accuracy",
"OOB accuracy",
"Test accuracy",
"Training sensitivity",
"OOB sensitivity",
"Test sensitivity",
"Training specificity",
"OOB specificity",
"Test specificity",
"Training balanced accuracy",
"OOB balanced accuracy",
"Test balanced accuracy",
"Training precision",
"OOB precision",
"Test precision",
),

```

```

Value = round(
  c(
    selected_trees,
    final_mtry,
    final_min_node_size,
    train_auc,
    oob_auc,
    test_auc,
    optimal_cutoff,
    train_accuracy,
    oob_accuracy,
    test_accuracy,
    train_sensitivity,
    oob_sensitivity,
    test_sensitivity,
    train_specificity,
    oob_specificity,
    test_specificity,
    train_balanced_accuracy,
    oob_balanced_accuracy,
    test_balanced_accuracy,
    train_precision,
    oob_precision,

```

```

        test_precision,
    ),
    3
)
)

cat(
  "final seed-123 model results\n"
)

cat(
  "Total wells :",
  nrow(model_data),
  "\n"
)

cat(
  "Training wells:",
  nrow(train_data),
  "\n"
)

cat(
  "Testing wells :",
  nrow(test_data),
  "\n\n"
)

print(summary_table)

cat(
  "\nFinal OOB Confusion Matrix\n"
)

print(
  oob_cm$table
)

cat(
  "\nFinal Test Confusion Matrix\n"
)

print(
  test_cm$table
)

```

```
cat(
  "\nFinal Training Confusion Matrix\n"
)
```

```
print(
  train_cm$table
)
```

```
cat(
  "\nPredictor Importance\n"
)
```

```
print(
  importance_df %>%
    select(
      Display_name,
      Importance
    )
)
```

```
write.csv(
  summary_table,
  file.path(
    out_dir,
    "model_summary.csv"
  ),
  row.names = FALSE
)
```

```
write.csv(
  importance_df,
  file.path(
    out_dir,
    "variable_importance.csv"
  ),
  row.names = FALSE
)
```

```
write.csv(
  test_data,
  file.path(
    out_dir,
    "test_predictions.csv"
  ),
)
```

```

    row.names = FALSE
  )

training_medians_df <- tibble(
  Predictor = names(training_medians),
  Training_median = as.numeric(training_medians)
)

write.csv(
  training_medians_df,
  file.path(
    out_dir,
    "training_medians.csv"
  ),
  row.names = FALSE
)

saveRDS(
  rf_model,
  file.path(
    out_dir,
    "final_random_forest_model.rds"
  )
)

capture.output(
  sessionInfo(),
  file = file.path(
    out_dir,
    "session_info.txt"
  )
)

library(ggplot2)

# Create figure folder
figure_folder <- file.path(
  out_dir,
  "Figures"
)

if (!dir.exists(figure_folder)) {
  dir.create(

```

```

    figure_folder,
    recursive = TRUE
  )
}

importance_df <- data.frame(
  Variable = names(rf_model$variable.importance),
  Importance = as.numeric(rf_model$variable.importance)
)

importance_df <- importance_df[
  order(
    importance_df$Importance,
    decreasing = TRUE
  ),
]

number_of_variables <- 18

importance_df <- head(
  importance_df,
  number_of_variables
)

variable_labels <- c(
  som = "Soil organic matter",
  o2_mg_l = "Dissolved oxygen",
  p_h = "pH",
  cultivated_10 = "Cultivated land",
  herbaceous_10 = "Herbaceous land",
  forest_10 = "Forest",
  herbaceous_wetlands_10 = "Herbaceous wetlands",
  satthick_2023_2025 = "Saturated thickness",
  predev_dtw_updated = "Predevelopment depth to water",
  dtw_2023_2025 = "Current depth to water",
  ksat = "Hydraulic conductivity",
  avg_pprec = "Average precipitation",
  corn_10 = "Corn",
  barren_10 = "Barren land",
  sum_nfert = "Nitrogen fertilizer",
  urban_m = "Distance to urban area",
  kdau_1 = "Feedlot density",
  all_streams_m = "Distance to streams"
)

importance_df$Label <- variable_labels[

```

```

importance_df$Variable
]

importance_df$Label[
  is.na(importance_df$Label)
] <- importance_df$Variable[
  is.na(importance_df$Label)
]

importance_df$Label <- factor(
  importance_df$Label,
  levels = rev(importance_df$Label)
)

importance_plot <- ggplot(
  importance_df,
  aes(
    x = Importance,
    y = Label
  )
) +
  geom_col(width = 0.7) +
  labs(
    x = "Permutation Importance",
    y = NULL
  ) +
  theme_classic(base_size = 12) +
  theme(
    text = element_text(
      family = "Times New Roman"
    ),
    axis.text = element_text(
      color = "black",
      size = 10
    ),
    axis.title = element_text(
      size = 11
    )
  )

print(importance_plot)

importance_plot_file <- file.path(
  figure_folder,
  "RF_variable_importance_top14.png"
)

```

```

ggsave(
  filename = importance_plot_file,
  plot = importance_plot,
  width = 7,
  height = 5,
  dpi = 600,
  bg = "white"
)

importance_csv_file <- file.path(
  figure_folder,
  "RF_variable_importance_top14.csv"
)

write.csv(
  importance_df,
  importance_csv_file,
  row.names = FALSE
)

cat("\nVariable importance figure saved to:\n")
cat(importance_plot_file, "\n")

cat("\nVariable importance table saved to:\n")
cat(importance_csv_file, "\n")


library(terra)
library(ranger)

raster_folder <- "D:/StatAnalysis/RFTS2/raster"

map_folder <- file.path(
  out_dir,
  "maps_30m"
)

temp_folder <- "D:/terra_temp"

if (!dir.exists(map_folder)) {
  dir.create(
    map_folder,

```

```

    recursive = TRUE
  )
}

if (!dir.exists(temp_folder)) {

  dir.create(
    temp_folder,
    recursive = TRUE
  )
}

terraOptions(
  tempdir = temp_folder,
  memfrac = 0.6,
  progress = 1
)

predictors <- c(
  "som",
  "o2_mg_l",
  "p_h",
  "cultivated_10",
  "herbaceous_10",
  "forest_10",
  "herbaceous_wetlands_10",
  "satthick_2023_2025",
  "predev_dtw_updated",
  "dtw_2023_2025",
  "ksat",
  "avg_pprec",
  "corn_10",
  "barren_10",
  "sum_nfert",
  "urban_m",
  "kdau_1",
  "all_streams_m"
)

# Raster file paths

raster_files <- c(

```

```

som = file.path(
  raster_folder,
  "som.tif"
),

o2_mg_1 = file.path(
  raster_folder,
  "o2_mg_1.tif"
),

p_h = file.path(
  raster_folder,
  "p_h.tif"
),

cultivated_10 = file.path(
  raster_folder,
  "cultivated_10.tif"
),

herbaceous_10 = file.path(
  raster_folder,
  "herbaceous_10.tif"
),

forest_10 = file.path(
  raster_folder,
  "forest_10.tif"
),

herbaceous_wetlands_10 = file.path(
  raster_folder,
  "herbaceous_wetlands_10.tif"
),

satthick_2023_2025 = file.path(
  raster_folder,
  "satthick_2023_2025.tif"
),

predev_dtw_updated = file.path(
  raster_folder,
  "predev_dtw_updated.tif"
),

dtw_2023_2025 = file.path(

```

```

    raster_folder,
    "dtw_2023_2025.tif"
  ),

  ksat = file.path(
    raster_folder,
    "ksat.tif"
  ),

  avg_pprec = file.path(
    raster_folder,
    "avg_pprec.tif"
  ),

  corn_10 = file.path(
    raster_folder,
    "corn_10.tif"
  ),

  barren_10 = file.path(
    raster_folder,
    "barren_10.tif"
  ),

  sum_nfert = file.path(
    raster_folder,
    "sum_nfert.tif"
  ),

  urban_m = file.path(
    raster_folder,
    "urban_m.tif"
  ),

  kdau_1 = file.path(
    raster_folder,
    "kdau_1.tif"
  ),

  all_streams_m = file.path(
    raster_folder,
    "all_streams_m.tif"
  )
)

file_check <- data.frame(

```

```

    predictor = names(raster_files),
    file = raster_files,
    exists = file.exists(raster_files)
)

print(file_check)

missing_files <- raster_files[
  !file.exists(raster_files)
]

if (length(missing_files) > 0) {

  print(missing_files)

  stop(
    "Some predictor raster files were not found. Check the filenames."
  )
}

cat(
  "\nAll predictor raster files were found.\n"
)

study_area_file <- file.path(
  raster_folder,
  "study_area.shp"
)

if (!file.exists(study_area_file)) {

  stop(
    "Study-area shapefile was not found. Check study_area_file."
  )
}

study_area <- vect(
  study_area_file
)

```

```
target_crs <- "EPSG:26914"
```

```
study_area <- project(  
  study_area,  
  target_crs  
)
```

```
o2_extent_raster <- rast(  
  raster_files[["o2_mg_1"]]  
)
```

```
if (nlyr(o2_extent_raster) > 1) {
```

```
  o2_extent_raster <- o2_extent_raster[[1]]  
}
```

```
# Project O2 to the target CRS
```

```
o2_extent_raster <- project(  
  o2_extent_raster,  
  target_crs,  
  method = "bilinear"  
)
```

```
cat(  
  "\nOriginal study-area extent:\n"  
)
```

```
print(  
  ext(study_area)  
)
```

```
cat(  
  "\nO2 raster extent:\n"  
)
```

```
print(  
  ext(o2_extent_raster)
```

```

    ext(o2_extent_raster)
)

mapping_extent <- intersect(
  ext(study_area),
  ext(o2_extent_raster)
)

if (is.null(mapping_extent)) {
  stop(
    "The study area and O2 raster do not overlap."
  )
}

template_30m <- rast(
  mapping_extent,
  resolution = c(30, 30),
  crs = target_crs
)

names(template_30m) <- "template"

print(template_30m)

cat(
  "\nTemplate resolution:",
  res(template_30m)[1],
  "x",
  res(template_30m)[2],
  "meters\n"
)

o2_mask_30m <- project(
  o2_extent_raster,
  template_30m,
  method = "bilinear"
)

```

```

# Restrict the O2 mask to the study-area polygon

o2_mask_30m <- mask(
  o2_mask_30m,
  study_area
)

# Convert valid O2 cells to 1
o2_mask_30m <- ifel(
  is.na(o2_mask_30m),
  NA,
  1
)

names(o2_mask_30m) <- "o2_coverage_mask"

print(o2_mask_30m)

plot(
  o2_mask_30m,
  main = "Final Mapping Area Based on O2 Coverage"
)

if (!exists("training_medians")) {
  stop(
    "Training medians were not found. Run the model section first."
  )
}

print(training_medians)

if (any(!is.finite(training_medians))) {
  stop(
    "At least one training median is NA."
  )
}

```

```

aligned_files <- character(
  length(predictors)
)

names(aligned_files) <- predictors

for (variable in predictors) {

  cat(
    "\nProcessing:",
    variable,
    "\n"
  )

  original_raster <- rast(
    raster_files[[variable]]
  )

  if (nlyr(original_raster) > 1) {

    original_raster <- original_raster[[1]]
  }

  # Align the raster to the 30 m O2-based template
  aligned_raster <- project(
    original_raster,
    template_30m,
    method = "bilinear"
  )

  median_value <- as.numeric(
    training_medians[[variable]]
  )

  # Replace missing values inside the template with the training median

  aligned_raster <- ifel(
    is.na(aligned_raster),

```

```

    median_value,
    aligned_raster
)

# Restrict to the study-area polygon
aligned_raster <- mask(
  aligned_raster,
  study_area
)

# Restrict further to the valid O2 coverage
aligned_raster <- mask(
  aligned_raster,
  o2_mask_30m
)

names(aligned_raster) <- variable

output_file <- file.path(
  map_folder,
  paste0(
    variable,
    "_30m.tif"
  )
)

writeRaster(
  aligned_raster,
  filename = output_file,
  overwrite = TRUE,
  datatype = "FLT4S",
  gdal = c(
    "COMPRESS=LZW",
    "TILED=YES"
  )
)

aligned_files[[variable]] <- output_file

```

```

rm(
  original_raster,
  aligned_raster
)

gc()
}

# Build the predictor stack from disk

predictor_stack <- rast(
  aligned_files[predictors]
)

names(predictor_stack) <- predictors

print(predictor_stack)

cat(
  "\nPredictor stack names:\n"
)

print(
  names(predictor_stack)
)

missing_predictors <- setdiff(
  predictors,
  names(predictor_stack)
)

if (length(missing_predictors) > 0) {

  print(missing_predictors)

  stop(
    "Some model predictors are missing from the raster stack."
  )
}

```

```

predictor_stack <- predictor_stack[[predictors]]

if (!identical(
  names(predictor_stack),
  predictors
)) {

  stop(
    "Raster names or order do not match the model predictors."
  )
}

geometry_check <- compareGeom(
  predictor_stack[[1]],
  predictor_stack[[2:nlyr(predictor_stack)]],
  stopOnError = FALSE
)

print(geometry_check)

if (!all(geometry_check)) {

  stop(
    "The predictor raster geometries do not match."
  )
}

cat(
  "\nPredictor stack created successfully.\n"
)

predict_high_probability <- function(
  model,
  data,
  ...
) {

```

```

prediction <- stats::predict(
  model,
  data = data
)$predictions

if (is.matrix(prediction)) {

  if ("High" %in% colnames(prediction)) {

    return(
      prediction[, "High"]
    )

  } else {

    return(
      prediction[, 2]
    )
  }
}

return(prediction)
}

#Create continuous probability map

probability_file <- file.path(
  map_folder,
  "nitrate_high_probability_30m.tif"
)

nitrate_probability <- terra::predict(
  predictor_stack,
  rf_model,
  fun = predict_high_probability,
  na.rm = TRUE,
  cores = 1,
  filename = probability_file,
  overwrite = TRUE,
  wopt = list(
    datatype = "FLT4S",
    gdal = c(
      "COMPRESS=LZW",

```

```

    "TILED=YES"
  )
)
)

names(nitrate_probability) <-
  "Probability_High_Nitrate"

print(nitrate_probability)

plot(
  nitrate_probability,
  main = "Probability of Nitrate >= 10 mg/L as N"
)

#Create and export binary low/high risk map

risk_cutoff <- optimal_cutoff

cat(
  "\nCutoff used for raster classification:",
  round(risk_cutoff, 3),
  "\n"
)

probability_for_risk <- rast(
  probability_file
)

# Create binary classes
nitrate_risk <- classify(
  probability_for_risk,
  rcl = matrix(
    c(
      -Inf, risk_cutoff, 0,
      risk_cutoff, Inf, 1
    ),
    ncol = 3,
    byrow = TRUE
  ),
  right = FALSE,

```

```

    include.lowest = TRUE
  )

  names(nitrate_risk) <- "Nitrate_Risk"

  risk_file <- file.path(
    map_folder,
    "nitrate_risk_binary_30m.tif"
  )

  writeRaster(
    nitrate_risk,
    filename = risk_file,
    overwrite = TRUE,
    datatype = "INT1U",
    NAflag = 255,
    gdal = c(
      "COMPRESS=LZW",
      "TILED=YES"
    )
  )

  risk_check <- rast(
    risk_file
  )

  cat(
    "\nExported binary raster frequencies:\n"
  )

  print(
    freq(risk_check)
  )

  cat(
    "\nExported binary raster range:\n"
  )

  print(

```

```

global(
  risk_check,
  "range",
  na.rm = TRUE
)
)

```

```

plot(
  risk_check,
  main = paste(
    "Exported Nitrate Risk Map\nCutoff =",
    risk_cutoff
  )
)

```

```

class_matrix <- matrix(
  c(
    0.00, 0.20, 1,
    0.20, 0.40, 2,
    0.40, 0.60, 3,
    0.60, 0.80, 4,
    0.80, 1.01, 5
  ),
  ncol = 3,
  byrow = TRUE
)

```

```

probability_classes <- classify(
  nitrate_probability,
  rcl = class_matrix,
  include.lowest = TRUE,
  right = FALSE
)

```

```

names(probability_classes) <-
  "Probability_Class"

```

```

class_file <- file.path(
  map_folder,
  "nitrate_probability_classes_30m.tif"
)

```

```

writeRaster(
  probability_classes,
  filename = class_file,
  overwrite = TRUE,
  datatype = "INT1U",
  gdal = c(
    "COMPRESS=LZW",
    "TILED=YES"
  )
)

```

```

plot(
  probability_classes,
  main = "Nitrate Probability Classes"
)

```

```

cat(
  "\nProbability raster summary:\n"
)

```

```

print(
  global(
    nitrate_probability,
    fun = c(
      "min",
      "mean",
      "max"
    ),
    na.rm = TRUE
  )
)

```

```

cat(
  "\nProbability raster range:\n"
)

```

```

print(
  global(
    nitrate_probability,
    "range",
    na.rm = TRUE
  )
)

```

```

    )
  )

cat(
  "\nBinary risk-cell counts:\n"
)

print(
  freq(nitrate_risk)
)

cat(
  "\nProbability-class cell counts:\n"
)

print(
  freq(probability_classes)
)

check_all_predictors <- function(
  train_data,
  predictor_stack,
  rf_model
) {

  vars <- rf_model$forest$independent.variable.names

  for (v in vars) {

    cat(
      "\nTraining:\n"
    )

    print(
      summary(
        train_data[[v]]
      )
    )
  }
}

```

```

cat(
  "\nRaster:\n"
)

print(
  global(
    predictor_stack[[v]],
    c(
      "min",
      "max",
      "mean",
      "sd"
    ),
    na.rm = TRUE
  )
)
}
}

check_all_predictors(
  train_data,
  predictor_stack,
  rf_model
)

cat(
  "\nRaster predictor names:\n"
)

print(
  names(predictor_stack)
)

cat(
  "\nRF model predictor names:\n"
)

print(
  rf_model$forest$independent.variable.names

```

```

)

cat(
  "\nRF probability columns:\n"
)

print(
  colnames(
    predict(
      rf_model,
      data = train_data
    )$predictions
  )
)

cat(
  "\nNumber of probability-raster cells:\n"
)

print(
  ncell(nitrate_probability)
)

```

Appendix B – Principal Component Analysis

```
library(tidyverse)
library(readxl)
library(factoextra)
library(e1071)
library(patchwork)
library(janitor)
library(missMDA)
library(psych)

input_file <- "D:/StatAnalysis/RFTS2/step8.xlsx"
out_dir <- "D:/StatAnalysis/RFTS2/PCA"

names(read_excel(input_file) %>% clean_names())

pca_variables <- c(
  "nitrate",
  "som",
  "cultivated_10",
  "herbaceous_10",
  "satthick_2023_2025",
  "ksat",
  "avg_pprec",
  "corn_10",
  "wte_2023_2025",
  "all_streams_m"
)

pca_data <- read_excel(input_file) %>%
  clean_names() %>%
  select(all_of(pca_variables)) %>%
  mutate(
    across(everything(), as.numeric)
  )

missing_summary <- data.frame(
  Variable = names(pca_data),
  Missing_Values = colSums(is.na(pca_data)),
  Missing_Percent = round(
    100 * colSums(is.na(pca_data)) / nrow(pca_data),
    2
  )
)
```

```

print(missing_summary)

cat(
  "Number of variables with missing values:",
  sum(missing_summary$Missing_Values > 0),
  "\n"
)

cat(
  "Number of observations with at least one missing value:",
  sum(!complete.cases(pca_data)),
  "\n"
)

cat(
  "Total missing cells:",
  sum(is.na(pca_data)),
  "\n"
)

write_csv(
  missing_summary,
  file.path(out_dir, "missing_value_summary.csv")
)

# Prepare complete data for KMO and Bartlett tests

pca_complete <- pca_data %>%
  filter(if_all(everything(), ~ !is.na(.) & is.finite(.)))

cat("\nRows before removing incomplete observations:",
    nrow(pca_data), "\n")

cat("Rows used for KMO and Bartlett tests:",
    nrow(pca_complete), "\n")

# Correlation matrix
cor_matrix <- cor(
  pca_complete,
  use = "complete.obs",
  method = "pearson"
)

# KMO Test

```

```

kmo_result <- psych::KMO(cor_matrix)

cat("\nOverall KMO:\n")
print(round(kmo_result$MSA, 3))

cat("\nIndividual KMO values:\n")
print(round(kmo_result$MSAi, 3))

# Bartlett's Test of Sphericity

bartlett_result <- psych::cortest.bartlett(
  cor_matrix,
  n = nrow(pca_complete)
)

cat("\nBartlett's Test of Sphericity:\n")
print(bartlett_result)

#skewness before transformation

skewness_before <- sapply(
  pca_data,
  function(x) {
    e1071::skewness(
      x,
      na.rm = TRUE,
      type = 2
    )
  }
)

cat("\nSkewness before transformation:\n")
print(round(skewness_before, 3))

#Transformation before imputation

pca_data <- pca_data %>%
  mutate(
    nitrate = log1p(nitrate),
    all_streams_m = log1p(all_streams_m),
    dtw_2023_2025 = sqrt(dtw_2023_2025),
    corn_10 = sqrt(corn_10),
  )

```

```

cat(
  "\nMissing values after transformation:",
  sum(is.na(pca_data)),
  "\n"
)

#Check skewness after transformation

skewness_after_transformation <- sapply(
  pca_data,
  function(x) {
    e1071::skewness(
      x,
      na.rm = TRUE,
      type = 2
    )
  }
)

cat("\nSkewness after transformation:\n")
print(round(skewness_after_transformation, 3))

# Remove observations with missing values
n_before <- nrow(pca_data)

pca_data <- pca_data %>%
  drop_na()

n_after <- nrow(pca_data)

cat(
  "\nNumber of observations before removing missing values:",
  n_before,
  "\n"
)

cat(
  "Number of observations after removing missing values:",
  n_after,
  "\n"
)

cat(
  "Number of observations removed:",

```

```

n_before - n_after,
"\n"
)

cat(
  "Missing values remaining:",
  sum(is.na(pca_data)),
  "\n"
)

# Check skewness after removing missing values

skewness_final <- sapply(
  pca_data,
  function(x) {
    e1071::skewness(
      x,
      type = 2
    )
  }
)

cat("\nFinal skewness values:\n")
print(round(skewness_final, 3))

# Save complete dataset used for PCA
write_csv(
  pca_data,
  file.path(out_dir, "PCA_complete_cases.csv")
)

pca_data <- pca_data %>%
  rename(
    `Nitrate` = nitrate,
    `Cultivated land` = cultivated_10,
    `Herbaceous land` = herbaceous_10,
    `Saturated thickness` = satthick_2023_2025,
    `Hydraulic conductivity` = ksat,
    `Average precipitation` = avg_pprec,
    `Corn` = corn_10,
    `Water table elevation` = wte_2023_2025,
    `Distance to streams` = all_streams_m,
    `Soil organic matter` = som
  )

# Check names

```

```

print(names(pca_data))

# Run PCA

pca_results <- prcomp(
  pca_data,
  center = TRUE,
  scale. = TRUE
)

# PCA summary
summary(pca_results)

# PCA loadings
round(pca_results$rotation, 3)

# PCA scores
head(pca_results$x)

# Convert PCA results to data frames
pca_scores <- as.data.frame(pca_results$x)
pca_loadings <- as.data.frame(pca_results$rotation)

p1 <- fviz_eig(
  pca_results,
  addlabels = TRUE
)

ggsave(
  file.path(out_dir, "scree_plot.png"),
  p1,
  height = 4,
  width = 6,
  dpi = 300
)

p2 <- fviz_pca_biplot(
  pca_results,
  label = "all"
)

ggsave(
  file.path(out_dir, "biplot.png"),
  p2,
  height = 6,
  width = 6,

```

```

    dpi = 300
  )

p3 <- fviz_pca_biplot(
  pca_results,
  axes = c(1, 2),
  label = "var",
  repel = TRUE
)

ggsave(
  file.path(out_dir, "loading1_2.png"),
  p3,
  height = 6,
  width = 6,
  dpi = 300
)

p4 <- fviz_pca_biplot(
  pca_results,
  axes = c(2, 3),
  label = "var",
  repel = TRUE
)

ggsave(
  file.path(out_dir, "loading2_3.png"),
  p4,
  height = 6,
  width = 6,
  dpi = 300
)

p5 <- fviz_pca_biplot(
  pca_results,
  axes = c(3, 4),
  label = "var",
  repel = TRUE,
  geom.ind = "none",
  col.var = "black"
)

ggsave(
  file.path(out_dir, "loading3_4_tst.png"),
  p5,
  height = 6,

```

```

width = 6,
dpi = 300
)

p6 <- fviz_contrib(
  pca_results,
  choice = "var",
  axes = 1
)

ggsave(
  file.path(out_dir, "contribution1.png"),
  p6,
  height = 6,
  width = 6,
  dpi = 300
)

p7 <- fviz_contrib(
  pca_results,
  choice = "var",
  axes = 2
)

ggsave(
  file.path(out_dir, "contribution2.png"),
  p7,
  height = 6,
  width = 6,
  dpi = 300
)

p8 <- fviz_contrib(
  pca_results,
  choice = "var",
  axes = 3
)

ggsave(
  file.path(out_dir, "contribution3.png"),
  p8,
  height = 6,
  width = 6,
  dpi = 300
)

```

```

p10 <- fviz_contrib(
  pca_results,
  choice = "var",
  axes = 4
)

ggsave(
  file.path(out_dir, "contribution4.png"),
  p10,
  height = 6,
  width = 6,
  dpi = 300
)

```

```

p9 <- fviz_pca_var(
  pca_results,
  axes = c(1, 2),
  col.var = "black",
  repel = TRUE
) +
  theme_minimal()

```

```

ggsave(
  file.path(out_dir, "circle1_2.png"),
  p9,
  height = 6,
  width = 6,
  dpi = 300
)

```

```

p92 <- fviz_pca_var(
  pca_results,
  axes = c(3, 4),
  col.var = "black",
  repel = TRUE
) +
  theme_minimal()

```

```

ggsave(
  file.path(out_dir, "circle3_4.png"),
  p92,
  height = 6,
  width = 6,
  dpi = 300
)

```

```

#color_plots

plot_1_2 <- fviz_pca_var(
  pca_results,
  axes = c(1, 2),
  col.var = "contrib",
  repel = TRUE
) +
  scale_color_gradientn(
    colours = c(
      "#2C7BB6",
      "#66C2A5",
      "#FDAE61",
      "#D7191C"
    ),
    limits = c(0, 25),
    breaks = c(10, 15, 20)
  ) +
  labs(
    color = "Contribution"
  ) +
  theme_minimal()

ggsave(
  file.path(out_dir, "Color1_2.png"),
  plot_1_2,
  height = 6,
  width = 6,
  dpi = 300
)

plot_3_4 <- fviz_pca_var(
  pca_results,
  axes = c(3, 4),
  col.var = "contrib",
  repel = TRUE
) +
  scale_color_gradientn(
    colours = c(
      "#2C7BB6",
      "#66C2A5",
      "#FDAE61",
      "#D7191C"
    ),
    limits = c(0, 40),

```

```

    breaks = c(10, 20, 30)
  ) +
  labs(
    color = "Contribution"
  ) +
  theme_minimal()

ggsave(
  file.path(out_dir, "Color3_4.png"),
  plot_3_4,
  height = 6,
  width = 6,
  dpi = 300
)

combined_plot <- plot_1_2 + plot_3_4

ggsave(
  filename = file.path(out_dir, "PCA_combined_plot.png"),
  plot = combined_plot,
  width = 14,
  height = 8,
  dpi = 300
)

contrib <- get_pca_var(pca_results)$contrib

make_contrib_plot <- function(pc_number) {

  plot_data <- data.frame(
    Variable = rownames(contrib),
    Contribution = contrib[, pc_number]
  )

  avg_contrib <- 100 / nrow(plot_data)

  ggplot(
    plot_data,
    aes(
      x = reorder(Variable, -Contribution),
      y = Contribution
    )
  ) +
  geom_col(fill = "steelblue") +
  geom_hline(
    yintercept = avg_contrib,

```

```

    linetype = "dashed",
    color = "red",
    linewidth = 0.8
  ) +
  labs(
    title = paste0(
      "Contribution of variables to Dim-",
      pc_number
    ),
    x = "",
    y = "Contributions (%)"
  ) +
  theme_minimal(base_size = 16) +
  theme(
    axis.text.x = element_text(
      angle = 45,
      hjust = 1
    ),
    plot.title = element_text(size = 20),
    axis.title.y = element_text(size = 16)
  )
}

# PC3
p_pc3 <- make_contrib_plot(3)

ggsave(
  file.path(out_dir, "contribution_PC3_fviz_style.png"),
  p_pc3,
  height = 6,
  width = 6,
  dpi = 300
)

# PC4
p_pc4 <- make_contrib_plot(4)

ggsave(
  file.path(out_dir, "contribution_PC4_fviz_style.png"),
  p_pc4,
  height = 6,
  width = 6,
  dpi = 300
)

library(flextable)

```

```

library(officer)
library(magick)

loadings <- round(
  pca_results$rotation[, 1:4],
  2
)

contributions <- round(
  get_pca_var(pca_results)$contrib[, 1:4],
  2
)

pca_table <- data.frame(
  Variable = rownames(loadings),
  Weight_PC1 = loadings[, 1],
  Weight_PC2 = loadings[, 2],
  Weight_PC3 = loadings[, 3],
  Weight_PC4 = loadings[, 4],
  Contribution_PC1 = contributions[, 1],
  Contribution_PC2 = contributions[, 2],
  Contribution_PC3 = contributions[, 3],
  Contribution_PC4 = contributions[, 4]
)

ft <- flextable(pca_table)

ft <- add_header_row(
  ft,
  values = c(
    "",
    "PCA Loadings",
    "Contribution to Variance (%)"
  ),
  colwidths = c(1, 4, 4)
)

ft <- set_header_labels(
  ft,
  Variable = "Variables",
  Weight_PC1 = "PC1",
  Weight_PC2 = "PC2",
  Weight_PC3 = "PC3",
  Weight_PC4 = "PC4",
  Contribution_PC1 = "PC1",
  Contribution_PC2 = "PC2",

```

```

    Contribution_PC3 = "PC3",
    Contribution_PC4 = "PC4"
)

black_line <- fp_border(
  color = "black",
  width = 1
)

ft <- border_remove(ft)

ft <- hline_top(
  ft,
  part = "header",
  border = black_line
)

ft <- hline(
  ft,
  i = 1,
  part = "header",
  border = black_line
)

ft <- hline_bottom(
  ft,
  part = "header",
  border = black_line
)

ft <- hline_bottom(
  ft,
  part = "body",
  border = black_line
)

ft <- flextable::font(
  ft,
  fontname = "Times New Roman",
  part = "all"
)

ft <- fontsize(
  ft,
  size = 10,
  part = "all"
)

```

```

)

ft <- align(
  ft,
  align = "center",
  part = "all"
)

ft <- align(
  ft,
  j = 1,
  align = "left",
  part = "all"
)

ft <- padding(
  ft,
  padding = 3,
  part = "all"
)

ft <- autofit(ft)

ft

save_as_image(
  ft,
  path = file.path(out_dir, "PCA_Table.png"),
  zoom = 3
)

img <- image_read(
  file.path(out_dir, "PCA_Table.png")
)

img <- image_background(
  img,
  color = "white",
  flatten = TRUE
)

image_write(
  img,
  path = file.path(out_dir, "PCA_Table.jpg"),
  format = "jpeg",
  quality = 100

```

)