Science of the Total Environment 957 (2024) 177418

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Mixed method analysis of anthropogenic groundwater contamination of drinking water sources in Malawi

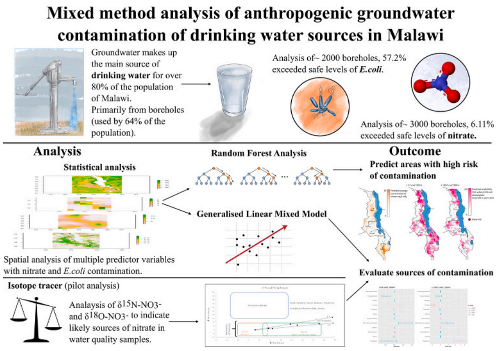
Rebekah G.K. Hinton [a,b,*], Robert M. Kalin [a], Limbikani C. Banda [a,c], Modesta B. Kanjaye [d], Christopher J.A. Macleod [b], Mads Troldborg [b], Peaches Phiri [c], Sydney Kamtukule [c]

[a] Department of Civil and Environmental Engineering, University of Strathclyde, Glasgow G1 1XJ, UK
[b] The James Hutton Institute, Craigiebuckler, Aberdeen AB15 8QH, UK
[c] Department of Water Resources, Ministry of Water and Sanitation, Government of Malawi, Private Bag 390, Lilongwe, Malawi
[d] Department of Sanitation and Hygiene, Ministry of Water and Sanitation, Government of Malawi, Private Bag 390, Lilongwe, Malawi

HIGHLIGHTS

- Large scale evaluation of nitrate and *E. coli* contamination in Malawi
- Sanitation variables most significant drivers of microbial and nitrate contamination
- Pilot isotope tracer analysis: sanitation as major source of nitrate contamination
- In particular, high densities of pit latrines result in groundwater contamination.
- Contamination a particular concern around peri-urban areas

GRAPHICAL ABSTRACT


Mixed method analysis of anthropogenic groundwater contamination of drinking water sources in Malawi
Groundwater makes up the main source of drinking water for over 80% of the population of Malawi.
Primarily from boreholes (used by 64% of the population).
Analysis of~ 2000 boreholes, 57.2% exceeded safe levels of E.coli.
Analysis of~ 3000 boreholes, 6.11% exceeded safe levels of nitrate.
Analysis
Statistical analysis
Spatial analysis of multiple predictor variables with nitrate and E.coli contamination.
Random Forest Analysis
Generalised Linear Mixed Model
Outcome
Predict areas with high risk of contamination
Evaluate sources of contamination
Isotope tracer (pilot analysis)
Analaysis of δ15N-NO3- and δ18O-NO3- to indicate likely sources of nitrate in water quality samples.

ARTICLE INFO

Editor: Jurgen Mahlknecht

Keywords:
Groundwater contamination
E. coli
Nitrate
Isotope tracing
Generalised linear model
Random forest
Sanitation

ABSTRACT

Groundwater contamination poses significant challenges to public health and sustainable development in Malawi, where approximately 80 % of the population relies on groundwater sources for drinking water. This study investigates the presence and drivers of nitrate and *E. coli* contamination in groundwater used for drinking. Analysis was conducted on results from 3388 boreholes/tube wells for nitrate contamination and 2418 drinking water sources drawn from groundwater for *E. coli* contamination. Overall, 6.11 % and 57.2 % of water-points did not meet WHO guidelines for safe drinking water quality for nitrate and *E. coli* contamination, respectively. Through a mixed-method approach, utilizing generalised linear mixed models and random forest regression modelling, the study identifies factors relating to sanitation usage as critical drivers of both nitrate and *E coli* contamination. Pit-latrine usage was identified as a particularly important factor in contamination; accounting for pit latrine density specifically, rather than population density, resulted in better model prediction for both nitrate and high *E. coli* contamination indicating that consideration of the specific type of sanitation is important in predicting water quality. In addition, a stable isotope tracer analysis method to validate predictions and monitor nitrate in drinking water was piloted, identifying human waste as a likely source of nitrate

[*] Corresponding author at: Department of Civil and Environmental Engineering, University of Strathclyde, Glasgow G1 1XJ, UK.
E-mail address: Rebekah.hinton@strath.ac.uk (R.G.K. Hinton).

https://doi.org/10.1016/j.scitotenv.2024.177418
Received 12 April 2024; Received in revised form 4 November 2024; Accepted 4 November 2024
Available online 16 November 2024

contamination. Overall, this study underscores the urgency of addressing sanitation-related contamination of drinking water sources to ensure access to safe drinking water in low-income settings.

1. Introduction

Groundwater is a critical resource, supplying safe and accessible drinking water for over 2 billion people internationally (Kundzewicz and Döll, 2009). Globally, 1.23 million deaths are attributed to unsafe water resources each year, with the burden of unsafe water twice as high in low-income countries (IHME, 2019). As a result, groundwater contamination poses an important concern for human health (Karunanidhi et al., 2021). Alongside human health repercussions, safeguarding groundwater quality is essential for environmental and ecosystem preservation (Li et al., 2021).

Contaminants are commonly categorised as deriving from natural or anthropogenic sources (Li et al., 2021). Anthropogenic sources of contamination, including agriculture and domestic wastewater (Li et al., 2021), pose a particular concern as population growth, urbanisation, industrialisation, and agricultural intensification are resulting in increasing levels (Li et al., 2021). As contamination from anthropogenic sources continues to accelerate, protecting groundwater is becoming an ever-more pressing issue.

Nitrate is one contaminant of concern related to human activities. Whilst nitrate does naturally occur in the environment as part of the nitrogen cycle, anthropogenic sources, predominantly from agriculture and domestic wastewater but also from mining activity, are major causes of nitrate contamination in groundwater (Morrissy et al., 2021; Harper et al., 2017; Morin and Hutt, 2009). Alongside being an ecological contaminant of concern, particularly in contributing to eutrophication of water systems (Nyenje et al., 2010), high nitrate levels have been associated with health risks including increased infant methemoglobinemia ‘blue baby’ syndrome as well as some cancers (Puckett et al., 2011; Rahman et al., 2021).

Emphasis is often placed on agricultural sources for nitrate contamination, with high nitrogen fertiliser application rates resulting in diffuse nitrate contamination of groundwater resources (Harper et al., 2017; Wick et al., 2012). However, other sectors can be as significant on water contamination. An analysis of the sources of nitrate contamination across Africa found that population density was a better indicator of groundwater nitrate contamination than fertiliser application on a continental level, with a lack of sanitation hypothesised to be the cause of elevated contamination in areas of high population density (Ouedraogo et al., 2019).The assessment of groundwater nitrate loads from human waste sources is thus essential in consideration of water contamination.

Alongside nitrate contamination, inadequate sanitation and domestic wastewater management has also been identified as a critical driver of microbial groundwater contamination (Back et al., 2018). A lack of sanitation infrastructure, resulting in open defecation, has been linked to contamination of groundwater used for drinking water in Asia and Africa (Kayembe et al., 2018; Okullo et al., 2017). However, poor sanitation can itself also cause groundwater contamination where wastewater is inappropriately discharged or leaked (Sridhar and Parimalarenganayaki, 2024) or where there is direct contamination from the sanitation infrastructure itself. Pit-latrines provide one example of how sanitation can result in direct microbial groundwater contamination. Serving as the primary source of sanitation for 1.8 billion people globally, pit-latrines are an integral component of sanitation internationally (Gwenzi et al., 2023). However, unless safely managed, pit-latrines can result in groundwater contamination (Banks et al., 2007; Chidavaenzi et al., 2000; Dzwairo et al., 2006; Escamilla et al., 2013; Graham and Polizzotto, 2013; Gwenzi et al., 2023; Islam et al., 2016; Ndoziya et al., 2019; Tillett, 2013; Wright et al., 2013), a particular concern when they are used in contexts with a high reliance on groundwater sources of drinking water (Graham and Polizzotto, 2013).

As well as contributing anthropogenic \ contaminants to water systems, human activities can also play a part in contamination pathways by influencing other components of the water cycle, such as climate-change related rainfall intensity (Adhikari and Nejadhashemi, 2016). In areas of high fertiliser application, heavy rainfall can result in groundwater contamination of nitrate through the leaching of nitrate from fertiliser (Bijay-Singh and Craswell, 2021). Heavy rainfall can also result in increased surface runoff, which may be contaminated with waste from pit latrines or open defecation, which can infiltrate boreholes and result in heightened microbial contamination (Aralu et al., 2022). In addition, the increased water table height following heavy rain can result in greater pit latrine effluent leaching into groundwater and contamination of boreholes (Rivett et al., 2022). Not only does this highlight the significance of a system's environmental context on anthropogenic contamination, but presents a potential growing challenge due to increased extreme weather events under climate change. *Building resilience to climatic extremes will require greater understanding of groundwater quality from human activities accounting for both sources of contamination and consequences of hydroclimatic extremes.*

Malawi represents a particularly pertinent case study in the consideration of groundwater quality management with one of the lowest levels of access to safe drinking water globally (UNICEF and WHO, 2024). Within Malawi, groundwater provides the main source of drinking water for almost 80 % of the population (NSO, 2021), making groundwater quality essential to providing safe drinking water provision. Boreholes/tubewells provide the main points of access of drinking water and range from under 10 m to over 60 m deep, with most boreholes 40-50 m deep (Kalin et al., 2019). Contamination is a major barrier to safe drinking water access; currently over 60 % of the population access drinking water from contaminated drinking water sources (NSO, 2021). Poor quality water infrastructure worsens the contamination crisis; the infiltration of polluted surface run-off into boreholes is increased in damaged or poor-quality boreholes due to cracks in the concrete apron (Rivett et al., 2022). This is particularly concerning in Malawi due to high rates of borehole non-functionality and minimal borehole maintenance (Truslove et al., 2019, 2020; Kalin et al., 2019) placing water infrastructure itself at a greater risk of contamination.

The health consequences of groundwater contamination at drinking water sources are exacerbated by a low level of water treatment, which is practised by under 40 % of the population (NSO, 2021). Even where water treatment is conducted, it is largely through inefficient treatment processes such as bleach chlorination (Nielsen et al., 2022). This makes microbial contamination of groundwater sources likely to result in direct consumption of contaminated drinking water by the 60 % of the population accessing water from contaminated sources (NSO, 2021). As such, inadequate groundwater quality is undermining Malawi's aim to provide 100 % of the population with clean water sources by 2030 (NPC, 2021).

The consequences of high levels of contamination of drinking water sources can be seen in the scale of waterborne disease challenges within Malawi, estimated to account for over half of the national disease burden (Chavula, 2021). Malawi's deadliest cholera outbreak on record occurred from 2022 to 2023 and was reported to be partially due to widespread drinking water contamination (Freeman et al., 2024). There have been growing concerns of faecal groundwater contamination from pit latrines as a factor in the high burden of waterborne disease (Pritchard et al., 2007, 2008). This is likely to worsen as, under current population growth scenarios, there is projected to be a three-fold increase in the number of water-points at high risk of pit latrine contamination due to proximity (Hinton et al., 2024b).

Yet despite the gross burden of groundwater contamination in

Malawi, no national level evaluation of the extent and sources of groundwater contaminants has been conducted. Previous studies exploring groundwater contamination in Malawi have been limited to sub-national scales (Rivett et al., 2022; Addison et al., 2020; Back et al., 2018; Pritchard et al., 2007, 2008; Mussa and Kamoto, 2023; Dzinjalamala et al., 2024; Mkandawire, 2008) and often do not empirically evaluate the sources of contamination (Pritchard et al., 2007, 2008; Mussa and Kamoto, 2023; Dzinjalamala et al., 2024; Mkandawire, 2008). This limits current understanding on the extent, spatial distribution, and origins of contaminated groundwater, restricting appropriate protection measures of groundwater. *National level analysis of groundwater contamination is needed to develop a representative picture of the scale of challenge of groundwater contamination, studies that explore not only the level of sources but apply various methods to investigate likely sources will be critical in developing appropriate management measures (Kalin et al., 2022a).*

Statistical models including Generalised Linear Mixed Modelling (GLMM) and Random Forest Regression (RF) can provide insight to the relationships between predictor variables and measured groundwater contamination. These models can be applied to enhance understanding of the sources of groundwater contamination (Ouedraogo et al., 2019) as well as predict areas likely to have high levels of contamination (Charulatha et al., 2017; He et al., 2022). Managing anthropogenic groundwater contamination requires both greater understanding of the sources of contamination as well as enhanced prediction of areas at risk of contamination requiring different statistical modelling techniques to achieve different goals. Both GLMM and RF models are particularly useful in their application to a broad range of data types and capacity to handle non-linear relationships (Liu, 2016; Louppe, 2014). GLMM models, alongside other linear regression models, have been used widely to explore sources of contamination of groundwater (Charulatha et al., 2017; Nolan and Hitt, 2006). They have benefit in robustly exploring the relationship between a response variable and predictor variables particularly as GLMMs can account for random effects as well as fixed effects (Rabe-Hesketh and Skrondal, 2008; Muschelli et al., 2014). As such they have been widely used to explain patterns in data in multiple fields (Goldstein and de Valpine, 2022; Zhu et al., 2007). However, as with all linear regression models, GLMM models are held back by their limited capacity to handle collinearity of variables (Hendrickx and Nutricia, 2018). This is a common challenge when investigating anthropogenic sources where multiple variables, e.g. population density and sanitation usage, are highly correlated, reducing model efficiency and making them less useful for accurate prediction.

RF models provide another tool to analyse and predict contamination trends. They have high predictive performance power (Couronné et al., 2018) particularly for spatial data (Hengl et al., 2018). The RF model functions as a combination of multiple decision trees with each tree applying a different subset of predictor variables to predict the response variable of the training dataset (Rokach and Maimon, 2015; Nath et al., 2022). They are particularly useful for collinear variables (Louppe, 2014). Whilst RFs indicate which predictor variables are most important in a specific model prediction (Ishwaran, 2007) variable importance must be interpreted with caution and cannot necessarily be used to indicate which are the most important predictor variables for the phenomena being studied (Louppe, 2014). As such, RFs have limited capacity in analysis of sources of contamination but are valuable for efficient prediction. In recognition of their specific strengths and limitations, combinations of GLMM and RF models have been utilised to enhance analysis and prediction (Bernaisch, 2022) and have been applied to studies of groundwater contamination (Ouedraogo et al., 2019; Charulatha et al., 2017; He et al., 2022; Nolan and Hitt, 2006).

Isotope hydrology is another commonly used method to evaluate groundwater contaminants and has been widely used for tracing sources of nitrate contamination which can be challenging to examine as they can be retained in groundwater for extended periods of time (Canter, 1996; Kendall et al., 2007; Jung et al., 2020, Nikolenko et al., 2018). By analysing the relative abundance of nitrogen and oxygen isotopes, likely sources can be identified due to characteristic patterns of isotope abundance, developing 'signatures' of the source of contamination. Whilst this method is highly effective, the need for specialised analytical facilities, not normally available in low-income countries, often makes application of the method non-feasible. In addition, this method has limited capacity to identify whether sources are from animal manure or human faecal waste due to their similar isotopic signatures (Kendall et al., 2007). *Applying mixed method analysis through the incorporation of statistical models alongside isotope hydrological analysis can enhance understanding of the extent and sources of groundwater contamination, developing understanding to enhance management approaches.*

This study adopts a mixed method analysis, using both GLMM and RF models to explore national groundwater contamination from multiple anthropogenic sources; both exploring sources of contamination (using GLMM) and predicting areas at high risk of contamination (using RF). Both methods are applied to two examples of contamination that are of concern in Malawi, microbial contamination (*E. coli* groundwater contamination) and nutrient contamination (nitrate contamination). Analysis is further enhanced through a pilot isotope analysis study of the sources of high nitrate levels. Specifically, this work addresses the following research questions and objectives: (1) What are the primary sources of nitrate and microbial groundwater contamination in Malawi? (2) What areas are predicted to have highest nitrate and microbial contamination? (3) Evaluate the use of stable isotopes of nitrate as a tool for nitrate source evaluation and verification of the model results in Malawi. These inferences provide valuable insight into groundwater management, informing decision making on contamination sources as well as identifying areas of concern for contamination and guiding areas for future water quality testing.

2. Methodology

2.1. Context and study area

Malawi is a country in South-Eastern Africa which experiences a tropical-continental climate, with a wet season from November to April and a dry season from May to October (Kalin et al., 2022a,b) (Fig. 1). Malawi's water supplies are dominated by Lake Malawi, both in the proportion of the country's water resources contained in Lake Malawi and in narratives around national water security and water resource management. In contrast, despite providing 82 % of water abstracted for agricultural, domestic, and industrial purposes, groundwater is an often overlooked facet of water security in Malawi (Fraser et al., 2020). Alongside being the biggest source of water for abstraction, groundwater underpins much of surface water security with baseflow from groundwater central to maintaining river flows (Kelly et al., 2020). This is particularly true for the dry season where over 90 % of all river flow comes from groundwater discharge (Kelly et al., 2020). Agricultural intensification is impacting Malawi's land and water management. Currently, the majority (over 80 %) of the population are employed in rain-fed, subsistence farming (NPC, 2021), however, planned governmental economic and agricultural development involves increased irrigation and fertiliser usage (MAIWD, 2018), placing growing pressure on water resources, notably groundwater.

Malawi is undergoing rapid demographic change with its current population of 21 million (World bank, 2024) anticipated to reach almost 60 million by the end of the century (United Nations, 2024). Urbanisation is also resulting in dramatic demographic shifts, with the 84 % of the population currently residing in rural areas anticipated to reduce to 40 % by 2063 (NPC, 2021). Groundwater forms the main source of drinking water for 80 % of the population (Kalin et al., 2022a,b; NSO, 2021) with water from boreholes/tubewells providing the main source of drinking water and used by 64 % of the population (NSO, 2021). Access to safely managed drinking water, defined as an 'improved water source that is accessible on premises, available when needed, and free

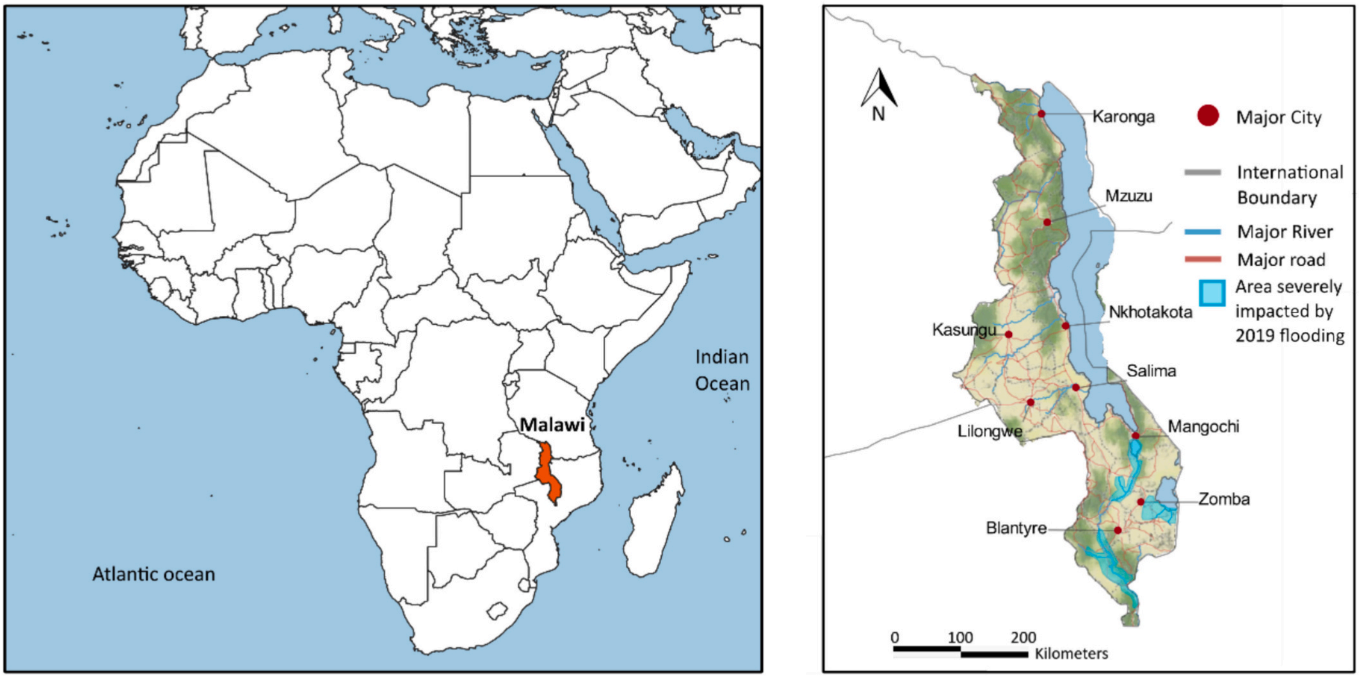


Fig. 1. Study location of Malawi showing major cities and rivers. Water quality data from across the country is analysed. Figure was produced in QGIS with Stamen Terrain background (QGIS, 2024).

from contamination' (WHO and UNICEF, 2017) is low; only 18 % of Malawi's population meet international guidelines for a safely managed drinking water source (UNICEF and WHO, 2024). Contamination is one of the major barriers to access of safely managed drinking water access with over 60 % of the population's source of drinking water having measurable *E. coli* contamination (NSO, 2021).

2.2. *Water quality data collection*

2.2.1. *Nitrate analysis*

Groundwater quality samples for nitrate analysis were collected from 3717 boreholes across Malawi. Samples were collected by Government of Malawi water laboratory staff through borehole construction contractors after drilling of a new borehole/tubewell. Information on borehole depth was not available for all boreholes but most water-points in Malawi have a depth between 40 and 50 m (Kalin et al., 2019). Samples were collected between 2000 and 2022, with most data collection from 2015 to 2022 due to an increased drilling effort. Data was provided by the Government of Malawi, Ministry of Water and Sanitation for this study. Out of the collected samples, 3388 were chosen for analysis after removing duplicate responses.

Water samples for nitrate analysis were collected in polyethylene bottles that were rinsed with distilled water, un-acidified, and stored at 4 °C during transportation to the government water laboratory in Malawi. The water samples were filtered through 0.45 μm Whatman filters prior NO_3-N analysis and measured against known laboratory standards. Before 2019, the HACH Chromotropic acid method was used with a HACH spectrophotometer. After 2019, samples were analysed using Ion Chromatography (Ion Analyzer—Model: IA-300). The NO_3-N analysis was performed following the International Standard Methods (APHA et al., 2005), and the accuracy of the results was confirmed through a series of quality assurance and control procedures specified in the International Standard Methods (APHA et al., 2005). The threshold of 50 mg NO_3^-/l was considered high nitrate according to the Malawi Standard (MS733:2005) for drinking water from boreholes and protected shallow wells as well as the WHO guideline standards (MBS, 2017; WHO, 2017).

2.2.2. *E. coli analysis*

Water quality *E. coli* levels were obtained from the Multiple Indicator Cluster Surveys (MICS); a nationally representative survey between December 2019–August 2020 of 26,882 households in 1111 clusters. This survey was conducted by the Government of Malawi National Statistical Office in collaboration with UNICEF (UNICEF, 1995). The survey sample was based on the 2018 Population and housing Census designed to provide representative clusters across the country. These surveys gathered household level information on a range of topics relevant to child, maternal, and family well-being. In addition to household survey responses, the MICs survey conducted water quality testing data at households' sources of drinking water, evaluating *E. coli* levels in household and source drinking water (Bain et al., 2021; NSO, 2021). Water quality testing was conducted at 2810 waterpoints within the 1111 clusters nationally. Georeferenced water quality data was provided by the Malawi National Statistical Office, UNICEF Malawi, and Global MICs team. Following data cleaning, 2801 complete and unique datapoints were selected. In this work, we evaluate only water-points drawn from groundwater, including boreholes/tubewells, dug wells, and protected springs. In total, water quality assessments for 2418 water-points drawn from groundwater were analysed.

E. coli water quality analysis followed a protocol outlined in the 2016 MICS Water Quality Testing Manual (UNICEF, 2016). A 100 ml water sample was obtained at sources of drinking water reported to be used by households. Prior to collection from the source, water was flushed for 30 s. Water samples were collected in sterilized 'Whirl Pak Bags', the water sample was subsequently filtered through a filter which was placed on an agar growth medium and incubated for 24 to 48 h and bacterial colony growth counted and recorded (UNICEF, 2016). The number of *E. coli* in a 100 ml sample of water was evaluated with values between 0 and 100 *E. coli* recorded as the number of *E. coli* and values exceeding >100 *E. coli* classed listed as 101 *E. coli*/100 ml.

As values exceeding 100 *E. coli* were not quantified, for the purposes of this study, binary classifications of *E. coli* contamination of water-points were created with water-points classified as any *E. coli* contamination (>0 *E. coli*/100 ml) and cases of very high *E. coli* contamination (≥100 *E. coli*/100 ml). Whilst *E. coli* contamination indicators were available for both source and household drinking water, we consider only source contamination as household drinking water is also significantly influenced by post-abstraction behavioural patterns such as water collection and storage methods as well as the types of containers being used for collection (Wright et al., 2004) and as such are not representative of groundwater contamination itself.

2.3. Water quality data visualisation

For data visualisation, binary contamination data of the presence of nitrate and *E. coli* contamination was rasterized to 10 km resolution using the rasterize() function within the raster package, R (Hijmans, 2024). The percentage of surveys conducted within each cell that exceeded given thresholds of contamination were calculated and summarized.

2.4. Statistical model variable selection

A range of socioeconomic and biophysical 'core' variables were selected for analysis within statistical models of groundwater contamination. Variables were selected based on the variables analysed in published methods (Ouedraogo et al., 2019; He et al., 2022) or where the literature suggested that greater exploration into specific variables (e.g. sanitation infrastructure) was needed (Ouedraogo et al., 2019).

The selected variables are summarized in Supplementary Information Table 2. Summary plots of spatial data used are shown in Supplementary Information Figs. 2 and 3.

The spatial distribution of different types of sanitation was identified as an area of interest. The types of sanitary facility provision considered were pit-latrine use, flush toilet use, and open defecation (no facility) as these make up the majority of sanitary access (Hinton et al., 2023). National spatial data regarding the type of sanitation was only available for pit-latrine usage.

For flush latrine usage and open defecation, spatial sanitation use data was produced following the methodology outlined in (Hinton et al., 2024b). A high resolution, 100 m, gridded population distribution of Malawi obtained from WorldPop population distribution (Worldpop, 2024) was defined as rural or urban areas based on the urban fraction outlined in (Hurtt et al., 2011). The rural and urban population for each district was multiplied by the respective level of sanitary facility use (or open defecation) for rural and urban populations as outlined in the 2015/16 DHS survey (NSO and ICF, 2017). Population density (Worldpop, 2024) and deprivation index (CIESIN, 2022) were also included as socioeconomic variables.

Another variable of interest was flooding extent. The 2019 Cyclone Idai flood was taken as a flooding event of interest as it was representative of other flooding events observed and was close to when most analysed data was collected. Flooding data was generated as a binary raster of areas impacted by flooding in 2018–2019, corresponding to the years leading up to water quality survey sampling. The raster was created in QGIS (QGIS, 2024), creating a map of flooded areas as reported in flooding report survey data (DoDMA, 2019; Scottish Government, 2019) and informed by stakeholder engagement (personal communication). In addition to flooding, the overall level of precipitation, precipitation trend, was obtained from the RCMRD Geoportal (RCMRD, 2015b).

Individual livestock distribution data was available from the gridded Livestock of the World (GLW 3) database (Gilbert et al., 2018). Total livestock data was calculated by summing the quantity of sheep, cows, pigs, and goats as these are the major mammalian livestock cultivated in Malawi. Cropland data was obtained from the HarvestPortal Database (FAO/NASA, 2024) and spatial data for both fertiliser and manure application was also included (Potter et al., 2010, 2012). In addition, anthropogenic biome was included as a measure of 'wildness' (RCMRD, 2015a).

Water sample specific information including the type of water source (e.g. dug well, borehole/tubewell, or protected spring) and the date of collection (month and year) was also included in model generation, this was only available for water-samples for *E. coli* contamination data and was not given for nitrate contamination data (which only evaluated boreholes/tubewells).

2.5. Multiple linear regression model construction

This study employed generalised linear mixed model (GLMM) structures to explore the relationship between response and predictor variables, accommodating noncontinuous as well as continuous variables with both fixed and random effects (Liu, 2016; Rabe-Hesketh and Skrondal, 2008). Three models of contamination were developed using binary response variables, the response variables in the respective models were the presence of high nitrate, any *E. coli* presence, and high *E. coli*. In each model, the response variable was modelled as a binary variable of whether contamination passed given thresholds.

For NO_3, the threshold for 'high nitrate' was 50 mg/l, with values at or exceeding this considered as high contamination according to national and WHO guidelines (MBS, 2017; WHO, 2017). For *E. coli* contamination, two GLMM models were constructed. The first *E. coli* GLMM model considered the presence of any *E. coli* contamination, therefore exceeding WHO specifications of safe drinking water (UNICEF and WHO, 2024). The second model considered high *E. coli* contamination, exceeding 100 *E. coli*/100 ml and considered as a 'very high' risk (NSO, 2021). All GLMM models used logistic regression as they applied continuous and categorical variables to a binary predictor.

All models were produced using the feGLM function in the fixest package (Bergé et al., 2023) in R (R Core Team, 2023), as this enabled GLMM generation with and without fixed effects (Bergé, 2018). For NO_3 contamination, a GLMM model with no fixed effects was constructed with continuous and categorical predictor variables and a binary NO_3 response variable. For *E coli* contamination, consistent data was available on the water source type and date of collection and were used as fixed effects. Within the *E. coli* contamination GLMM water source and date (year and month) were included as fixed effects. The number of levels for each fixed effect is summarized in the model structure in the Supplementary Information, Tables 9 and 12. Both *E. coli* contamination models therefore had a binary contaminant response variable with categorical and continuous predictor variables with fixed and random effects.

GLMM probabilistic assumptions of linearity, response distribution, independence and multicollinearity were confirmed using the R functions lm and the Variance Inflation Factor (VIF) (Chambers, 1992; R Core Team, 2023; Wilkinson and Rogers, 1973). Diagnostic plots and VIF factors are provided in Supplementary Information, Figs. 4–6 and Tables 4–6.

Where there was high multicollinearity between the predictor variables, two GLMMs were generated for each contaminant model containing all variables without high multicollinearity as well as one of the identified variables with high multicollinearity. The model performance of the two GLMMs, each containing one of the highly collinear variables, was evaluated and the model with best overall performance, for each contaminant, is summarized within the results.

Data was subset into training and testing data, using 60 % training to 40 % testing data. The GLMM model was applied to predict testing data outcomes using the R predict function (R Core Team, 2023), predicted contamination was compared to measured data and a confusion matrix produced. Metrics for model evaluation are summarized in Eqs. (1)–(4). Model performance metrics are accuracy (proportion of cases correctly categorised) (Eq. 1), precision (proportion of positive cases identified) (Eq. 2), sensitivity (proportion of predicted positives that were true positives) (Eq. 3), and specificity (proportion of negatives that were true negatives) (Eq. 4). Model fit was also evaluated using the McFadden pseudo R^2 (McFadden, 1974) (Eq. 5), calculated within the feGLM function, fixest package (Bergé et al., 2023), R. For McFadden pseudo R^2

values between 0.2 and 0.4 represent an 'excellent fit'(McFadden, 1977).

$$Accuracy = (\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}) \tag{1}$$

$$Precision = \mathrm{TP}/(\mathrm{TP}+\mathrm{FP}) \tag{2}$$

$$Sensitivity = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \tag{3}$$

$$Specificity = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP}) \tag{4}$$

$$R^2_{McF} = 1 - ln(L_M)/ln(L_0) \tag{5}$$

where TP is true positive (the number of cases correctly predicted as positive), TN is true negative (the number of cases correctly predicted as negative), FP is false positive (the number of cases incorrectly predicted as positive), and FN is false negative (the number of cases incorrectly predicted as negative). R^2_{McF} is the McFadden pseudo R^2, L_M is the likelihood of the fitted model and L_0 is the likelihood of the null model.

For the models of NO_3 contamination and high *E. coli* contamination presence, there was imbalance in the dataset with only a small percentage of samples exceeding the given thresholds. To improve model development, the minority class was upsampled using the upsample() function under the caret package in R (Kuhn, 2008) to make class distributions equal.

2.6. Random Forest model construction

For spatial prediction of areas of contamination, random forest modelling was applied using the package randomForest in R (Breiman, 2001; Liaw and Wiener, 2002). The number of decision trees was set as 500, considered to be an appropriate balance to limit overfitting. All given spatial predictor variables were included in the model for both NO_3 and *E. coli* level. For NO_3 contamination, a continuous response variable for NO_3 was predicted, therefore a regression random forest model was generated. For *E. coli* contamination, two binary models of *E. coli* contamination were produced: presence of any *E. coli* contamination and *E. coli* contamination of 100 *E. coli*/100 ml and above. For the generation of these random forest models, unsupervised random forest models were produced.

Data was split into training and testing datasets, using 70 % for model training and 30 % for testing. Model performance was evaluated by calculating the Root Mean Square Error (RMSE), summarized in Eq. (6), and R^2 coefficient for the continuous, regression model, of NO_3, Eq. (7).

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2} \tag{7}$$

where N is the number of data points, y_i is the given value of y, $\hat{y}$ is the predicted value of y, and $\bar{y}$ is the mean value of y.

For categorical model prediction of *E. coli* contamination, model performance was evaluated using a confusion matrix to calculate model accuracy, precision, sensitivity, and specificity (Eqs. 1–4). Feature importance was evaluated by calculating Shapley values of variables in the Random Forest Model, comparing model predictions with and without features being included, model simulations were iteratively run to give different feature orders. Shapley values were calculated using a randomly assigned 5 % subset of the data due their computationally intensive nature. The kernelshap() function, within the kernelshap package in R (Mayer et al., 2023), was used to calculate Shapley values. Shapley values were visualized as a beeswarm plot using the shapviz() and sv_importance() functions within the shapviz package, R (Mayer and Stando, 2024).

For visualisation of predicted contamination, the random forest models were applied to create a raster of predicted contamination. The NO_3 model generated a predicted NO_3 concentration raster for the average level value of NO_3 contamination for a water-point within a given 10 km cell. For the *E. coli* models, the percent for waterpoints within a given 10 km cell that would exceed thresholds of *E. coli* was predicted. Predicted rasters were produced by applying the random forest model to a raster stack of all predictor variables using the predict() function under the raster package in R (Hijmans, 2024). Maps of predicted contamination were produced in QGIS for visualisation (QGIS, 2024).

2.7. Isotope analysis

A pilot study using nitrate isotope analysis was undertaken within the Linthipe river sub-catchment (Kalin et al., 2022b) in the central region of Malawi. The dominant aquifer type within the Linthipe sub-catchment is a colluvium overlying weathered and fractured basement (Kalin et al., 2022b) with extensive groundwater-surface water connections within the region. Pilot samples were collected as part of an International Atomic Energy Agency (IAEA) national project (MWL-7002 TC project). Targeted groundwater and surface water samples were collected at 15 locations suspected of high nitrate concentrations between May and June 2015 and shipped to the IAEA (Vienna) for analysis of δ^{15}N and δ^{18}O of NO_3^-.

Water samples were collected in 60 ml HDPE bottles tapped tightly to prevent evaporation and exchange with atmospheric water vapor and stored in cool conditions (4 °C) during transportation and holding at the Isotope Hydrology Laboratory of the IAEA (Vienna, Austria). The δ^{15}N-NO_3^- and δ^{18}O-NO_3^- were measured using dual isotope approach and results were reported in per mil (‰) relative to atmospheric air (N_2) and Vienna Standard Mean Ocean Water (VSMOW) standards for nitrogen and oxygen, respectively. International reference materials (IAEA-NO_3^-, USGS34 and USGS35) were used for data calibration and instrumental monitoring. Analytical precision was less than ±0.2 ‰ for δ_{15}N- NO_3^-, and ±0.5 ‰ for δ^{18}O-NO_3^-, respectively.

3. Results

3.1. Spatial distribution of groundwater contaminants

Fig. 2 summarises the percent of water quality surveys conducted within 10 × 10 km grid cells which surpassed water quality thresholds. A map of surveying intensity is provided in the Supplementary Information, Fig. 2.

Of the 3388 complete water quality tests surveying NO_3, 207 (6.11 %) exceeded the WHO threshold of 50 mg/l (WHO, 2017) with an average NO_3 level of 3.1 mg/l. There were 322 cases of contamination over 10 mg/l, exceeding historic Malawi Standards guidelines (Pullanikkatil et al., 2015; MBS, 2017), and 212 cases exceeding the current Malawi standards guidelines of 45 mg/l (Chidya et al., 2016; MBS, 3017). Overall, of the 2418 MICs water quality surveys, 1383 (57.2 %) water-points had *E. coli* contamination surpassing WHO guidelines of 0 *E. coli*/100 ml (NSO, 2021) and. 361 (14.9 %) water-points had 100 or more *E. coli*/100 ml.

3.2. Multiple linear regression contamination model selection

There was high multicollinearity between population density and (pit) latrine density for all models. To meet the assumption of collinearity, one of the variables with high multicollinearity (latrines and population) was removed. For NO_3, there was also high multicollinearity between the predictor variables flush toilet use and open defecation. To resolve this case, flush toilet use was removed, VIF values before and after the removal of flush toilet usage are shown in the

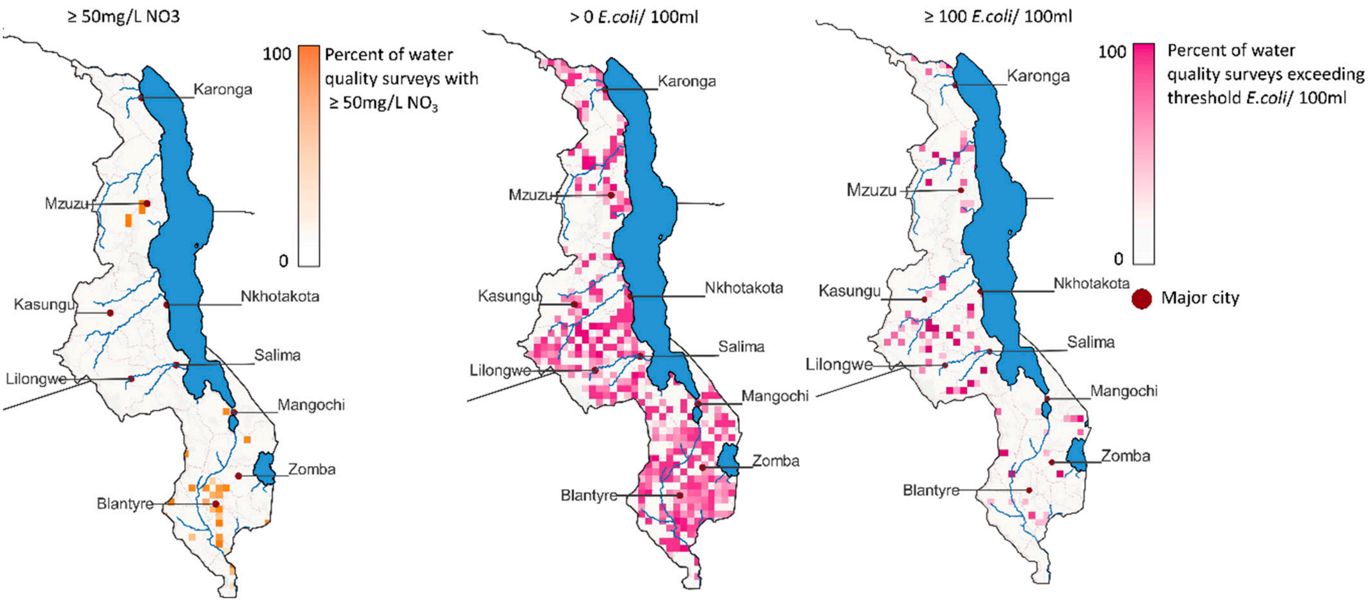


Fig. 2. The percent of groundwater samples within a 10 × 10 km grid exceeding thresholds of drinking water quality for nitrate and *E. coli*. Figure produced in QGIS (QGIS, 2024).

appendix. Flush toilet use did not have high multicollinearity in the *E. coli* contamination model and therefore was included. Following the removal of variables with high multi-collinearity, all VIF values were below 3 and met the assumption of collinearity (James et al., 2013). For each contamination model, models were produced for all variables excluding latrines and another model with all variables excluding population. Model fit was evaluated and is summarized in Table 1. The model with the highest model accuracy was selected as the GLMM model used for further analysis.

For nitrate above 50 mg/l and high *E. coli* contamination (≥100 *E. coli*/100 ml), the models including latrine density as a predictor variable had higher accuracy than the model including population density; nitrate dependent variable models had 70.92 % and 68.23 % accuracy where population and latrine density were dependent variables respectively, models with high *E. coli* as the dependent variable had 80.97 % and 78.52 % accuracy for models with population and latrines as dependent variables respectively. For *E. coli* presence (>0 *E. coli*/100 ml) the model with population density as a predictor variable resulted in higher accuracy (57.98 % accuracy) than the model with latrine density (56.02 % accuracy). All diagnostic plots and assumptions for the selected GLMM for each contaminant are provided in the Supplementary Information Tables 4–6 and Figs. 4–6.

3.3. NO_3 contamination GLMM

The binomial GLMM (with no fixed effects) for NO_3 contamination had good model performance with a McFadden Pseudo R^2 value of 0.329 (considered excellent fit), 70.9 % overall model accuracy, 50 % sensitivity and 71.3 % specificity.

Predictor variable estimates are presented in Fig. 3. Precipitation was the significant predictor variable with the highest estimate, with areas with higher precipitation reporting a lower chance of high NO_3 contamination (≥50 mg NO_3/l). A similar effect was also observed for flooding, areas with high flooding had a lower chance of having high NO_3. Areas with higher anthropogenic biome, a measure of 'wildness', with high anthropogenic biome values being further away from both urbanised areas and intensive cropping, had more NO_3 contamination. This was also seen in that cropping intensity, pit-latrine density, and livestock density were negatively correlated with the presence of high NO_3 contamination. Water points in areas with high poverty were also less likely to report high nitrate levels. Areas with a high catchment level density of pit-latrine users (WRU Latrine User Density) had more waterpoints with nitrate values exceeding safe limits. This was the only factor to have a significant positive correlation with the presence of high NO_3 contamination.

3.4. *E. coli* contamination GLMM

Two binomial GLMM (fixed effects of date and water source) were produced for *E. coli* contamination, the results are summarized in Fig. 4. For the presence of *E. coli* contamination (>0 *E. coli*/100 ml), the GLMM model had a McFadden R^2 value of 0.08 and 58.0 % accuracy indicating

Table 1
Model performance for the contaminant models containing either latrine density or population density. Multiple metrics are shown. The highest value for each metric is highlighted in bold. Core variables are Anthropogenic Biome, Cropping Intensity, Fertiliser, Flooding, Manure, Pit-Latrine Density, WRU Pit-latrine density, Livestock, Open Defecation, Poverty, and Precipitation. The fixed effects on sample source and month of collection were included for the *E. coli* models. McFadden Pseudo R^2 values between 0.2 and 0.4 are considered an excellent fit.

Contaminant model	Variables included in model	McFadden Pseudo R^2	Accuracy	Precision	Sensitivity	Specificity
≥50 mg NO_3/l	Core Variables + Population	0.306	**70.92 %**	3.209 %	50.00 %	**71.32 %**
	Core Variables + Pit-latrines	**0.329**	68.23 %	**3.421 %**	**54.17 %**	68.52 %
>0 *E. coli*/100 ml	**Core Variables + Fixed effects + Flush Toilet + Population**	**0.0790**	**57.98 %**	72.17 %	**57.58 %**	58.72 %
	Core Variables + Fixed effects + Flush Toilet + Pit-latrines	0.0775	56.02 %	**75.35 %**	47.93 %	**71.00 %**
≥100 *E. coli*/100 ml	Core Variables + Fixed effects + Flush Toilet + Population	0.2075	**80.97 %**	**41.67 %**	57.25 %	**85.31 %**
	Core Variables + Fixed effects + Flush Toilet + Pit-latrines	**0.2404**	78.52 %	39.07 %	**61.31 %**	81.70 %

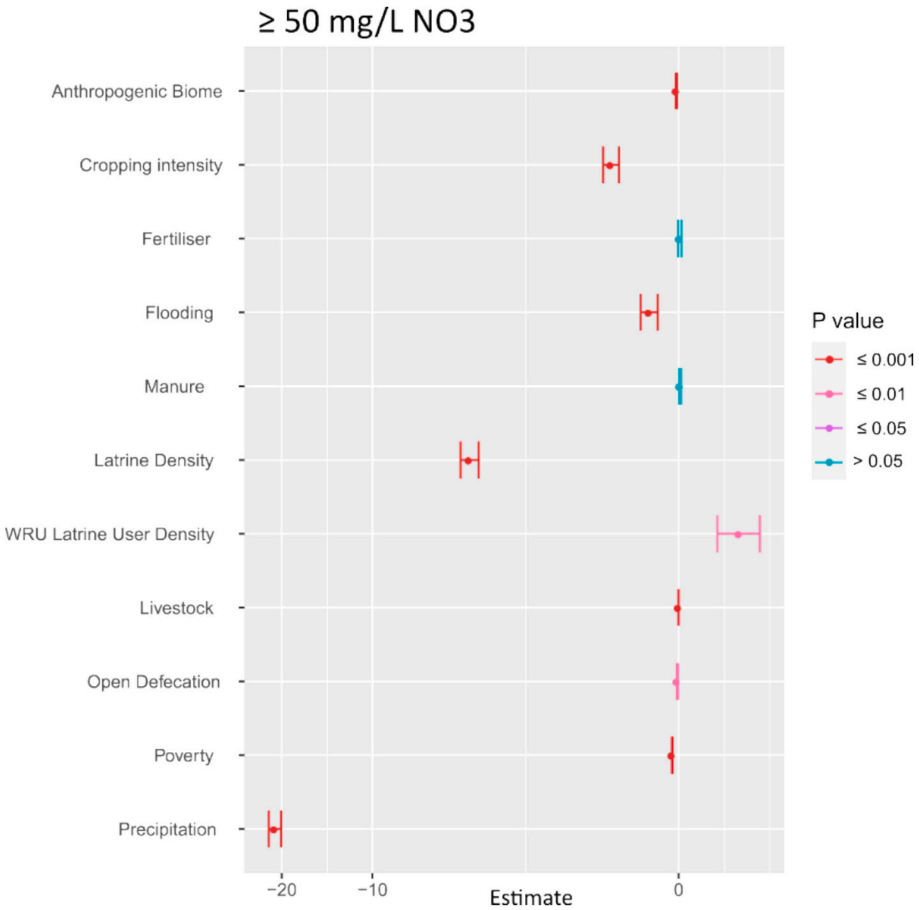


Fig. 3. GLMM model for the presence of high nitrate contamination (≥50 mg NO_3/L) of water-points. Coefficient estimate is shown on the x axis. Significant variables are highlighted in red and pink. Non-significant variables are shown in blue.

moderate performance. The area being impacted by 2019 flooding and the density of people practising open defecation were significantly correlated with an increased presence of *E. coli* contamination. Areas with a high density of flush toilet usage were significantly less likely to have some *E. coli* contamination.

For high *E. coli* contamination (≥100 *E. coli*/100 ml), the model had an 'excellent fit' with a McFadden R^2 of 0.24 and a high accuracy of 78.5 %. Precipitation and pit-latrine density significantly resulted in an increased risk of high *E. coli* contamination. Livestock density and the area being impacted by 2019 flooding were significantly negative drivers of high *E. coli.*

3.5. Random Forest prediction of contamination

A regression RF model for NO_3 contamination was generated for continuous data of NO_3 levels. Overall, the RF model had good model performance with a RMSE of 10.6 and a R^2 fit of 0.87. The plot of predicted vs measured nitrate contamination is shown in the Supplementary Information Fig. 8. Overall, the model underestimated nitrate contamination, particularly in cases where there was very high contamination.

As *E. coli* contamination was predicted as a binary variable (whether contamination was above set thresholds) the *E. coli* contamination RF model was evaluated by confusion matrix model performance metrics (Eqs. 1–4). For predictions of where there was some *E. coli* contamination, the random forest model had an average error rate of 30.0 % (70.0 % accuracy). The model performed better than the multiple linear regression model for all metrics.

For high *E. coli* contamination (≥100 *E. coli*/100 ml) the model had a 19.1 % error rate (81 % accuracy). The model performed better than the multiple linear regression model for accuracy and specificity, as it performed well at identifying areas without high *E. coli* contamination. However, the model failed to identify some of the cases of high contamination and performed worse for precision and sensitivity. Confusion matrices for both *E. coli* models are provided Supplementary Information Tables 7 and 9. The spatial distribution of areas of predicted contamination is summarized in Fig. 5.

The RF model of NO_3 contamination underpredicted some areas with high contamination but had good performance (R^2 0.87). Areas with predicted high levels of nitrate contamination were predominantly surrounding the cities of Blantyre and Mzuzu (Water Resource Areas 1 and 7). For the presence of some *E. coli* contamination, rural areas as well as areas in the north (surrounding Karonga), along the Shire River, and peri-urban areas outside of the major cities were predicted to have high levels of some *E. coli* contamination. The model of whether a water-point had ≥100 *E. coli*/100 ml predicted similar spatial distribution to the presence of any *E. coli* but with regions in the north (surrounding Karonga) as well as areas close to major cities, predicted to have high levels of high *E. coli* contamination. Areas of high NO_3 were typically surrounding population centres with NO_3 levels reducing further away from the population centre. In comparison, *E. coli* contamination was typically more restricted in location with a less clear gradient surrounding population centres and instead regions of high *E. coli* found generally along rounds and the outskirts of major population centres. This is likely due to *E. coli* contamination being highly localised to contamination sources as it cannot survive for extended periods of time in groundwater systems whilst nitrate contamination can travel large distances in groundwater (Canter, 1996).

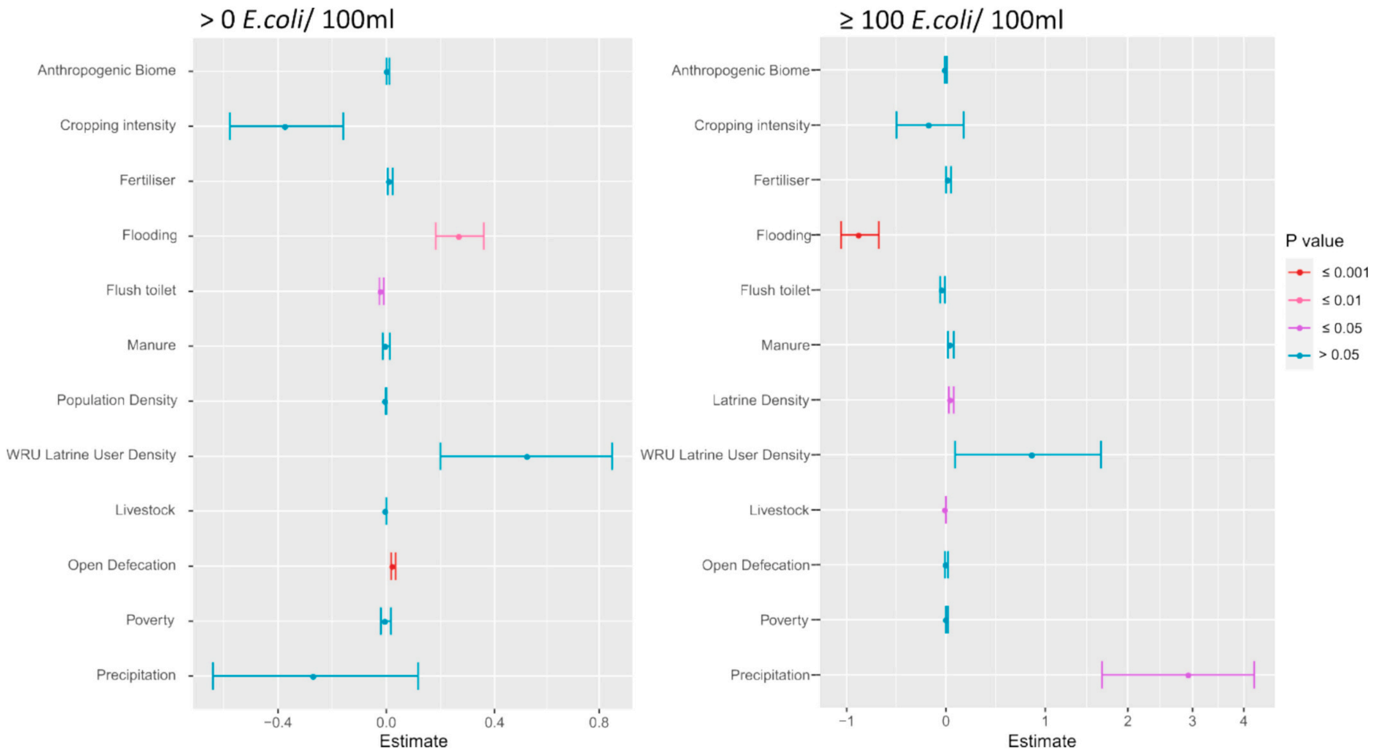


Fig. 4. Fixed effect GLMM models for the presence of any (>0 E coli/100 ml) and high contamination (≥100 *E. coli*/100 ml) *E. coli* of water-points. Coefficient estimate is shown on the x axis. Significant variables are highlighted in red and pink. Non-significant variables are shown in blue.

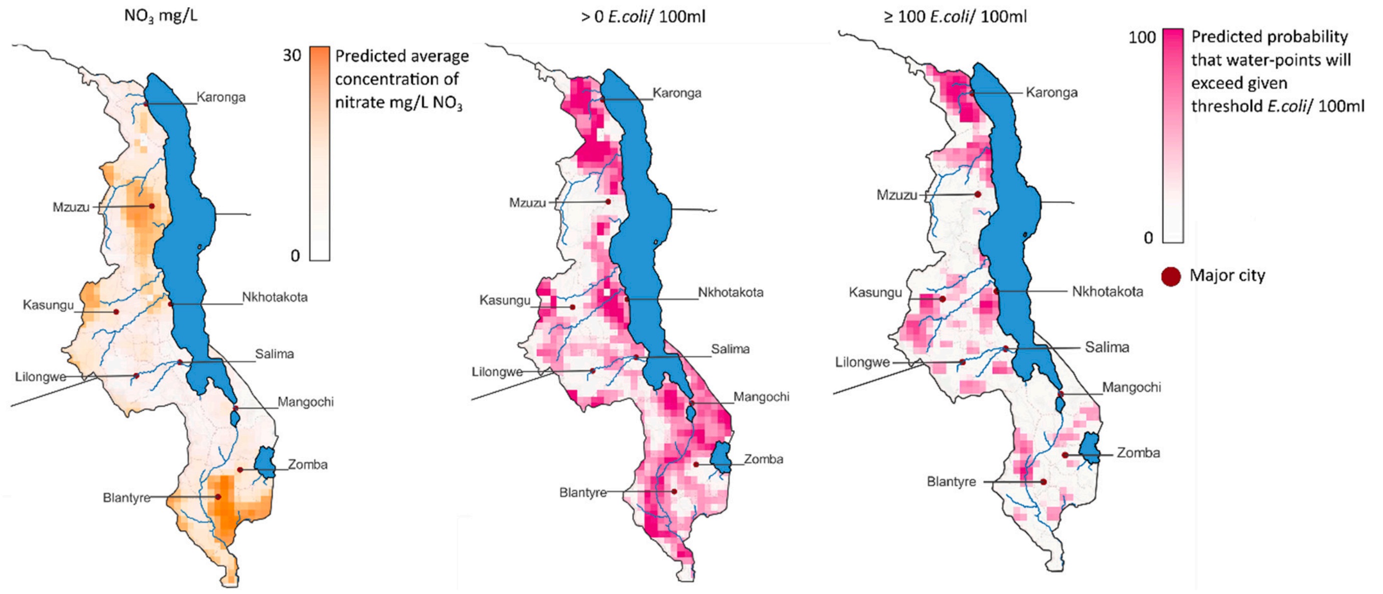

NO_3 mg/L
> 0 *E.coli*/ 100ml
≥ 100 *E.coli*/ 100ml
30
0
Predicted average concentration of nitrate mg/L NO_3
100
0
Predicted probability that water-points will exceed given threshold *E.coli*/ 100ml
Major city
Karonga
Mzuzu
Kasungu
Nkhotakota
Salima
Lilongwe
Mangochi
Zomba
Blantyre

Fig. 5. Predicted contamination from random forest models of NO_3 and *E. coli*. NO_3 contamination was modelled as a continuous response variable with the predicted amount of nitrate contamination given as NO_3 mg/l. The model underpredicted some of the areas with highest contamination but had overall good performance with an R^2 of 0.87. The presence of >0 *E. coli*/100 ml and ≥100 *E. coli*/100 ml were modelled as binary variables with discrete response variables. The model of whether a given water-point had >0 *E. coli*/100 ml had 70 % accuracy. The model of whether a water-point had ≥100 *E. coli*/100 ml had 81 % accuracy. Figure produced in QGIS (QGIS, 2024).

Fig. 6 shows Shapley values, showing the contribution of each feature within the random forest model produced for each contaminant considered. Variables are ordered by their mean feature Shapley value, indicating feature importance. Each point represents the influence of a variable on a given model simulation. For all cases, flooding was the feature with least contribution. For NO_3 and high *E. coli* contamination (≥100 *E. coli*/100 ml), sanitation related variables were the two variables with the highest contributions (WRU latrine user density and latrine density for NO_3 and flush toilet and open defecation for *E. coli*). For *E. coli* presence (>0 *E. coli*/100 ml), cropping and anthropogenic biome had the highest contribution.

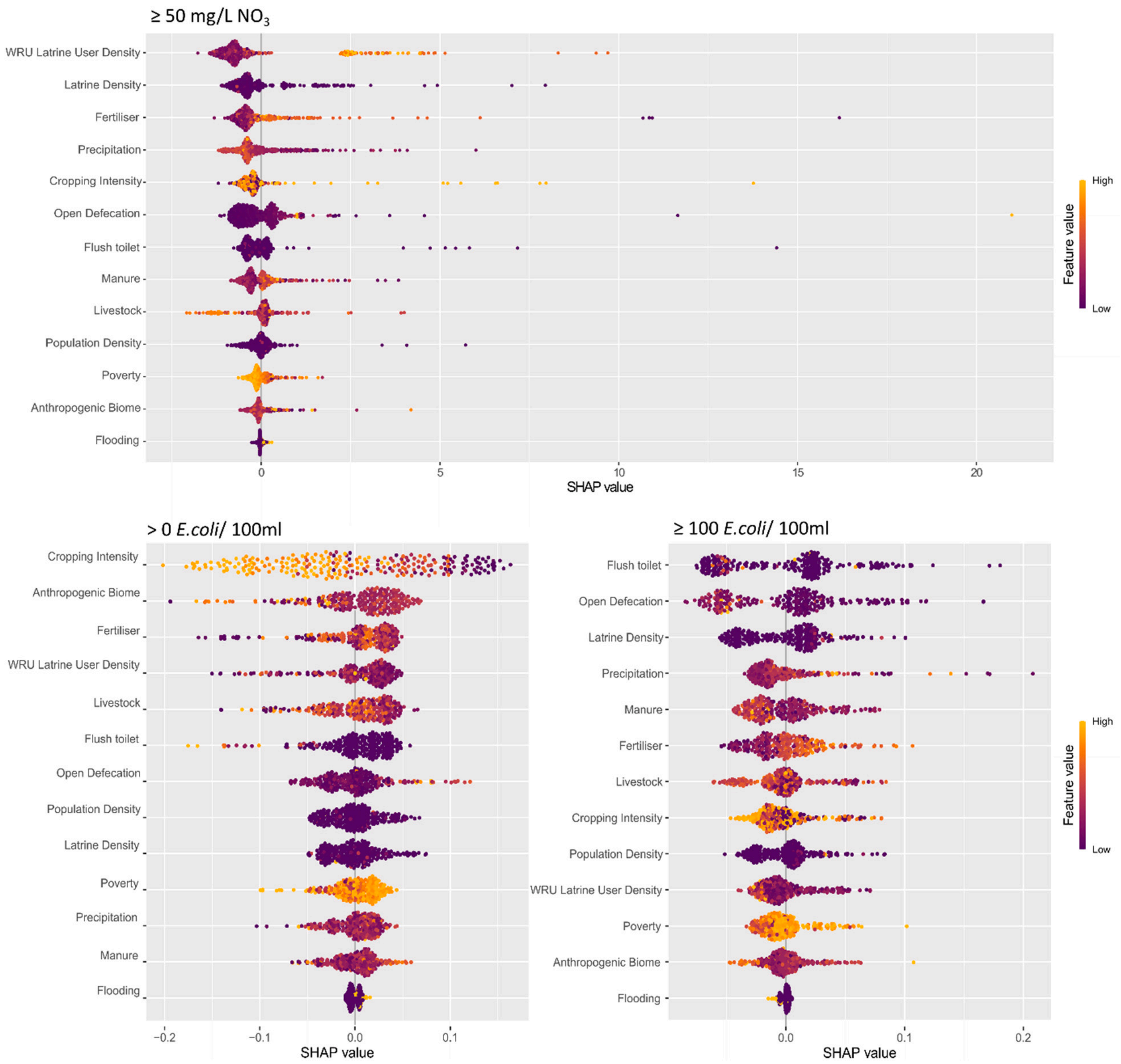


Fig. 6. Shapley values of variables within the Random Forest Regression model. The colour gives feature values with low feature values in blue/purple and high feature values in yellow/orange. SHAP values show the impacts of each feature on the model in each observation run, positive SHAP values indicate that the feature resulted in positive contribution to the chance of contamination whilst negative values respond to a negative contribution of the contamination. Features are ordered by the feature's mean Shapley value (an indicator of feature importance).

3.6. *Pilot isotope study*

Stable isotope hydrology was introduced in Malawi by the IAEA to enhance monitoring and management of water resources. This pilot study was part of a wider application of stable isotopes across the country (Banda et al., 2019; Banda et al., 2021; Banda et al., 2024). Stable isotopes of nitrate have the potential to validate the sources of nitrogen compounds in the water environment. Of the 15 groundwater and surface water samples collected within the pilot, 40 % (6) had concentrations of NO_3^-N at or above 0.1 mg/l concentration which warrants measurement of $\delta^{15}N\text{-}NO_3^-$ and $\delta^{18}O\text{-}NO_3^-$. Samples were analysed in triplicate, the $\delta^{15}N\text{-}NO_3^-$ ranged from −1.9 (±1.7) to 27.7 (±1.3)‰ with a mean of 11.4 (±0.6), whilst the $\delta^{18}O\text{-}NO_3^-$ ranged from 0.0 (±1.0) to 16.3 (±0.3)‰, with a mean of 8.6 (±1), values are shown in Supplementary Information Table 19.

Whilst the dataset is limited, the results hint that the most likely source of nitrate in surface water and groundwater originated as oxidised ammonia (NH_4^+), Fig. 7. The results also suggest that manure or human waste is a likely source of the ammonia (Kendall et al., 2007), Fig. 7. The dataset is not sufficient to track source terms and dynamics (Minet et al., 2017), but it does support the findings of this paper, pointing to pit-latrine derived nitrate in groundwater being a concern, and clearly shows a strong potential for further study to validate the predictions put forward in this paper.

4. Discussion

4.1. *Sources of contamination*

Nitrate and *E. coli* are two contaminants of concern for Malawi's water provision. Nitrate pollution is a public health concern and has been a growing concern in water quality in Malawi (Chidya et al., 2016; Chimphamba and Phiri, 2014; Nkwanda et al., 2021; Pullanikkatil et al., 2015; von Hellens, 2013.,) with high levels reported in both surface

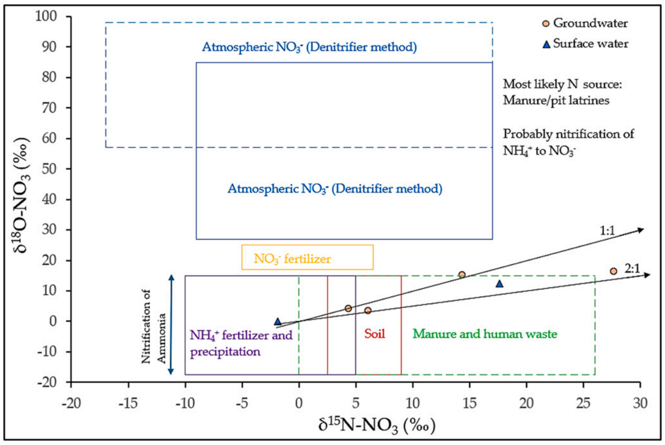


Fig. 7. Results of $\delta^{15}N\text{-}NO_3^-$ and $\delta^{18}O\text{-}NO_3^-$ plotted with the likely source of N species and trends added after Kendall et al., 2007.

water (Nkwanda et al., 2021; Pullanikkatil et al., 2015; Sajidu et al., 2007) and groundwater sources (Chidya et al., 2016; Chimphamba and Phiri, 2014; von Hellens, 2013). High groundwater nitrate pollution has been linked to contamination from sanitation sources, both within Malawi (von Hellens, 2013; Back et al., 2018) and beyond (Templeton et al., 2015; Ouedraogo et al., 2019; Rahman et al., 2021). Increasing loading of nitrate to groundwater is a particular concern for safeguarding water quality; nitrate does not undergo reduction in aerobic environments and therefore remains in groundwater for extended periods. When nitrate does not undergo reduction and remains in groundwater it can be transported over large distances, making sources of contamination hard to trace (Canter, 1996).

A GLMM was generated to explore sources of nitrate contamination of over 3000 water sources nationally with 70.9 % accuracy, catchment level density of pit-latrine usage (modelled density of pit-latrine users within WRU) was identified as the only significant positive driver of high nitrate levels. This suggests that there may be a high degree of transport of nitrate from domestic wastewater occurring in groundwater, at catchment levels, in Malawi. Areas with high densities of pit-latrines themselves were significantly less likely to have high levels of nitrate groundwater contamination. High pit-latrine density is mostly found in areas of high population density in which high concentrations of leachate will more likely result in anaerobic conditions in groundwater, enabling denitrification and therefore not resulting in high levels of nitrate. Pit-latrine related variables (WRU pit-latrine user density and pit-latrine density) were also identified as the features with the greatest contributions within the continuous RF model of nitrate contamination (R^2 0.87). As in the case of the GLMM, a high density of pit-latrine users per WRU increased the likelihood of high nitrate in groundwater whilst higher densities of pit-latrines themselves were negatively correlated with nitrate contamination within the model, although care should be taken in interpretation of RF variable importance. The relationship between population density/pit latrines and nitrate in groundwater was validated via the results of the pilot study of $\delta^{15}N\text{-}NO_3^-$ and $\delta^{18}O\text{-}NO_3^-$ in groundwater and surface water samples that indicates manure and/or pit-latrines present the major source of nitrate contamination.

Collectively considering these results, our findings strongly support previous examinations of nitrate contamination sources which identified sanitation sources as key drivers of nitrate pollution (Chidya et al., 2016; Chimphamba and Phiri, 2014; Ouedraogo et al., 2019; von Hellens, 2013). Ouedraogo et al., 2019, found that population density was a better predictor of pan-African nitrate levels than fertiliser, suggesting that the lack of sanitation in large areas of the African continent may be the reason for population density resulting in greater nitrate contamination (Ouedraogo et al., 2019). We build upon this inference, suggesting that pit-latrines themselves, more specifically pit-latrine density on a catchment/sub-catchment scale, is a major source of groundwater nitrate contamination. A nation-wide study that monitors nitrate and stable isotopes is recommended to monitor the projected growth of pit-latrine usage within Malawi (Hinton et al., 2023) and consequent growing risk of high densities of faecal waste loading (Hinton et al., 2024b).

In addition to anthropogenic factors, nitrate concentration was also significantly influenced by precipitation and flooding events, with lower nitrate concentrations reported in areas with high precipitation and being impacted by flooding, as shown in Fig. 3. This is likely due to high precipitation resulting in dilution of groundwater nitrate, as has been widely recorded (Wick et al., 2012; Nakagawa et al., 2021; Boumans and Fraters, 2010; Mas-Pla and Menció, 2019). Alternative studies have identified instances of high precipitation increasing nitrate concentrations where precipitation events have been suggested to enhance leaching of nitrate from soils (Liu et al., 2024), the absence of an apparent increase in soil nitrate leaching with precipitation perhaps further points to soil nitrate not being a major source of nitrate within this case study.

E. coli contamination is a significant barrier to achieving access to safe drinking water within Malawi (NSO, 2021; Mkandawire, 2008; Dzinjalamala et al., 2024; Mussa and Kamoto, 2023). Nationally, 57.2 % of drinking water sources drawn from groundwater were found to have some presence of *E. coli* and therefore not meet WHO guidelines for safe drinking water. Of particular concern were the 14.9 % of drinking water sources drawn from groundwater that show exceptionally high levels of *E. coli* contamination with 100 or more *E. coli* in a 100 ml sample. To evaluate drivers and provide spatial prediction of *E. coli* contamination in drinking water sources drawn from groundwater, we applied categorical GLMM and RFR models. We considered two cases of *E. coli* contamination, evaluating both the presence of any *E. coli* in drinking water (exceeding WHO guidelines) and another model evaluating very high levels of *E. coli* contamination ($\geq$100 *E. coli*/100 ml). Both cases were modelled as binary variables of the presence/absence of any/very high *E. coli* contamination.

For the presence of both any and high *E. coli* contamination, sanitation related variables were identified as critical drivers. A high density of people practising open defecation was a significant positive indicator of the presence of any *E. coli* contamination whilst flush toilet usage was negatively correlated with the presence of any *E. coli* within the GLMM. In areas where there is a high level of open defecation, environmental contamination because of open defecation may result in contamination of drinking water sources drawn from groundwater. Such environmental contamination by open defecation can result in contamination of groundwater water-points through contaminated surface water and runoff (Rivett et al., 2022). Water-point contamination from contaminated surface runoff alongside an elevated groundwater table promoting increased pit-latrine groundwater contamination has also been reported during flooding (Rivett et al., 2022). The geographic areas most impacted by heavy flooding in 2019 typically in the south of the country near the Shire River (approximately 1 year prior to water quality testing), were more likely to have evidence of *E. coli* contamination. Conversely, areas impacted by 2019 flooding, were significantly less likely to have very high *E. coli* contamination. This may be due to the flooded areas having been impacted by floods a year prior to the water quality tests being conducted and cases of exceptionally high *E. coli* may have undergone intervention over this time.

These findings underline the importance of community wide approaches in ending open defecation to prevent drinking water contamination (Hinton et al., 2024b). Unless safe sanitation for all is provided (as outlined in SDG 6.2), safe drinking water provision may be undermined. Water-points which are damaged or partially functional are more vulnerable to contamination from contaminated surface water (Rivett et al., 2022); a particular concern in Malawi due to limited maintenance and high non-functionality of water-points (Kalin et al., 2019; Kalin et al., 2022a,b). Combating groundwater *E. coli* contamination should involve not only sustainable progress towards ending open defecation but also ensuring improved borehole maintenance and functionality (Kalin et al., 2019) such as promoting community-led solutions to borehole functionality alongside ending open defecation (Hinton et al., 2021).

4.2. Predicted spatial distribution of contamination

Prediction of the distribution of nitrate and *E. coli* contamination using RFR models enabled greater spatial investigation of areas at high risk of contamination. Spatial prediction of areas susceptible to nitrate contamination identified water-points within water resource areas (WRA) 1 and 7, around the cities of Mzuzu and Blantyre, to be more likely to have high nitrate contamination. These areas have a high density of pit-latrine users within these catchments and also have limited precipitation. Spatial prediction of areas with any *E. coli* contamination (70 % accuracy) predicted that areas with any *E. coli* contamination were more likely to be in rural or peri-urban areas with a high density of people practising open defecation and susceptible to flooding. Spatial prediction of the highest contamination cases, where there were 100 or more *E. coli* per 100 ml, had 81 % accuracy. Cases of high contamination were mostly predicted in densely populated, non-urban areas with high pit-latrine density and high precipitation, typically in peri-urban towns or along roads.

4.3. Study limitations

Samples used for statistical analysis of nitrate contamination were gathered through the Government of Malawi when new boreholes were established. Overall, a national dataset of 3388 boreholes was analysed. To gather such an extensive dataset, samples collected over a 22-year period were analysed, although most samples were collected after 2015. For statistical analyses, spatial rasters of given predictor variables were used, these were typically circa 2020, although ranged from circa 2010. Selecting only samples taken within a smaller time window would have resulted in smaller sample sizes as well as samples that were not nationally representative thereby reducing the statistical power of the analysis. This limitation was deemed appropriate considering that spatial patterns of predictor variables were consistent over time. Future campaigns gathering national, extensive samples of nitrate contamination, as was seen for microbial contamination in the 2019/20 MICs survey, should be prioritised to enable increased analysis.

The level of *E. coli* contamination was provided as a binary variable for samples up to and including 100 *E. coli*/100 ml, however, when contamination was >100 *E. coli*/100 ml, this was marked as a binary measure. As such, for the purposes of this analysis, *E. coli* contamination was considered as a binary (presence of any *E. coli* contamination, and 100 or more *E. coli*/100 ml). This limitation was a result of the sampling method used for which it is hard to count >100 *E. coli* within a sample. This restricted analysis to binary methods or to only considering cases below 100 *E. coli*/100 ml. As this work was particularly interested in high contamination, binary analysis was completed. Further insights could be facilitated, including providing a better prediction of the level of *E. coli* contamination, through analyses with continuous variables. Considering the high number of water-points with exceptionally high *E. coli* contamination, alternative sampling methods enabling high *E. coli* concentrations to be measured, should be explored. The binary discretisation of the presence or absence of *E. coli* may also be a reason for the relatively poor model performance for the presence of any *E. coli* contamination.

Here we present a pilot study of isotopic analysis of nitrate contamination sources of a sample region in the Linthipe river subcatchment. Whilst the results do indicate nitrate sources from manure and domestic wastewater, only 6 samples underwent isotopic analysis. This limitation was due largely to resources, with the isotopic analysis unable to be conducted in Malawi and having to be conducted at the IAEA in Vienna, Austria. As such, only 15 samples were considered for isotopic analysis. Of those samples, only 6 had nitrate levels with sufficient nitrate to conduct the analysis. To not further restrict the results, 2 of the 6 samples were surface water, this was considered relevant for inclusion due to the high connectivity observed in Malawi between groundwater and surface water (Kelly et al., 2020).

4.4. Policy recommendations and future work

This study supports previous findings within Malawi, and on a continental scale, that sanitation infrastructure is a critical consideration for both nitrate (Templeton et al., 2015; Ouedraogo et al., 2019; Rahman et al., 2021) and microbial (Pritchard et al., 2007, 2008) groundwater contamination. The study emphasises the importance of community wide improvements in sanitation access as open defecation and poor sanitation infrastructure can result in contamination of community-based water resources. This echoes the ethos of programmes such as Community Led Total Sanitation (CLTS) which emphasise the environmental health component of enhanced sanitation provision (Chambers and Kar, 2008; Hinton et al., 2024a). However, whilst these initiatives push to end open defecation at the community level this work highlights the importance of community wide changes in sanitation not only focusing on eliminating open defecation but also on evaluating appropriate pit-latrine usage and management (Hinton et al., 2024a). *Ending open defecation on a community level is important but environmental health perspectives of inappropriate sanitation should also be emphasised.*

This work highlights a paradox in Malawi's progress in sanitation and water; open defecation must be eliminated to improve sanitation and water access but pit-latrines, which often form 'starter sanitation' (UNICEF, 2018) may cause contamination themselves unless appropriately managed. *Appropriate pit-latrine use will be important in ensuring an end to open defecation without resulting in widespread water contamination.* To ensure progress in both spheres of water and sanitation, enhanced policy frameworks to foster cooperation between stakeholders should be promoted. Sanitation infrastructure development must consider groundwater contamination consequences. Critically, this involves guiding long-term investment into higher quality waste management that minimizes contamination considering future projections of high population growth and increasing pit-latrine usage (Hinton et al., 2024b).

Our findings support initiatives to target water quality monitoring in areas of concern. Further expansion of isotope analysis may facilitate tracing of groundwater contaminant sources and develop evidence for source contamination management. Further understanding of contamination, both nitrate and microbial contamination, is needed to guide intervention. Understanding sanitation use will be critical (pit-latrine density was a better predictor of contamination of nitrate and high *E. coli* contamination than population density). Future work and modelling efforts should account for additional factors such as soil type (He et al., 2022) and permeability, as well as localised groundwater dynamics to enhance understanding of areas at high risk of contamination.

5. Conclusion

We apply a mixed method approach to identify drivers of microbial and nitrate contamination of groundwater drinking sources in Malawi. A pilot application of isotope hydrology was used to validate likely sources of nitrate contamination of groundwater. Statistical analysis was used to further enhance understanding of sources of nitrate contamination with catchment level pit-latrine usage identified as a significant driver of areas with high nitrate groundwater contamination. These findings support previous analyses of groundwater in Malawi and across Africa of sanitation sources being a major driver of groundwater nitrate contamination (Templeton et al., 2015; Ouedraogo et al., 2019; Rahman et al., 2021). Pit latrines were noted as a specific concern, highlighting the need for understanding of how sanitation derived contamination occurs. The results raise concerns for future groundwater contamination with projected increases of pit-latrine usage in Malawi driven by a move to end open defecation alongside high population growth (Hinton et al., 2021).

Sanitation related factors were significant considerations for microbial groundwater contamination. The density of open defecation was found to be a significant driver in cases of any *E. coli* contamination whilst in areas with very high *E. coli* contamination, pit-latrines are most likely the source of contamination. Policy and research efforts need to navigate how appropriate sanitation can be provided to ensure an end to open defecation without coming at the cost of groundwater quality. We also found that flooding risk is an important predictor of microbial borehole contamination, however, more research with higher quality flood risk predictive data would enhance understanding of the future risks of water-point contamination due to climate-change enhanced flooding.

CRediT authorship contribution statement

Rebekah G.K. Hinton: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Robert M. Kalin:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Methodology, Formal analysis, Conceptualization. **Limbikani C. Banda:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Modesta B. Kanjaye:** Writing – review & editing, Conceptualization. **Christopher J.A. Macleod:** Writing – review & editing, Conceptualization. **Mads Troldborg:** Writing – review & editing, Conceptualization. **Peaches Phiri:** Writing – review & editing, Conceptualization. **Sydney Kamtukule:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Robert Kalin reports financial support was provided by Scottish Government. Robert Kalin reports financial support was provided by International Atomic Energy Agency. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was conducted as part of a Scottish Government funded joint PhD studentship under the Climate Justice Fund Water Futures Program.

Georeferenced *E. coli* data was provided from the Malawi 2019/20 MICs Survey was provided by the UNICEF Malawi and the Global MICs team on behalf of Malawi National Statistical Office (MNSO). Nitrate borehole water quality samples were provided by the Government of Malawi, Ministry of Water and Sanitation. The pilot nitrate isotope tracing work was conducted as part of an International Atomic Energy Agency (IAEA) national project (MWL-7002 TC project).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.177418.

Data availability

Data will be made available on request.

References

Addison, M.J., Rivett, M.O., Robinson, H., Fraser, A., Miller, A.M., Phiri, P., Mleta, P., Kalin, R.M., 2020 Apr 10. Fluoride occurrence in the lower East African Rift System, Southern Malawi. Sci. Total Environ. 712, 136260. https://doi.org/10.1016/j.scitotenv.2019.136260. Epub 2019 Dec 26. PMID: 31945540.

Adhikari, U., Nejadhashemi, A.P., 2016. Impacts of climate change on water resources in Malawi. J. Hydrol. Eng. 21 (11). https://doi.org/10.1061/(asce)he.1943-5584.0001436.

APHA, AWWA, WEF, 2005. Standard Methods for the Examination of Water and Wastewater, 21st edition. American Public Health Association, American Water Works Association and Water Environment Federation, Washington (DC).

Aralu, C.C., Okoye, P.A.C., Abugu, H.O., Eze, V.C., 2022. Pollution and water quality index of boreholes within unlined waste dumpsite in Nnewi, Nigeria. Discov. Water 2. https://doi.org/10.1007/s43832-022-00023-9.

Back, J.O., Rivett, M.O., Hinz, L.B., Mackay, N., Wanangwa, G.J., Phiri, O.L., Songola, C. E., Thomas, M.A.S., Kumwenda, S., Nhlema, M., Miller, A.V.M., Kalin, R.M., 2018. Risk assessment to groundwater of pit latrine rural sanitation policy in developing country settings. Sci. Total Environ. 613–614, 592–610, 2018. ISSN 0048-9697.

Bain, R., Johnston, R., Khan, S., Hancioglu, A., Slaymaker, T., 2021. Monitoring drinking water quality in nationally representative household surveys in low-and middle-income countries: cross-sectional analysis of 27 multiple indicator cluster surveys 2014-2020. Environ. Health Perspect. 129. https://doi.org/10.1289/EHP8459.

Banda, L.C., Rivett, M.O., Kalin, R.M., Zavison, A.S.K., Phiri, P., Kelly, L., Chavula, G., Kapachika, C.C., Nkhata, M., Kamtukule, S., et al., 2019. Water–isotope capacity building and demonstration in a developing world context: isotopic baseline and conceptualization of a Lake Malawi Catchment. Water 11, 2600. https://doi.org/10.3390/w11122600.

Banda, L.C., Rivett, M.O., Zavison, A.S.K., Kamtukule, S., Kalin, R.M., 2021. National stable isotope baseline for precipitation in Malawi to underpin integrated water resources management. Water 13, 1927. https://doi.org/10.3390/w13141927.

Banda, L.C., Kalin, R.M., Phoenix, V., 2024. Isotope hydrology and hydrogeochemical signatures in the Lake Malawi Basin: a multi-tracer approach for groundwater resource conceptualisation. Water 16, 1587. https://doi.org/10.3390/w16111587.

Banks, D., Karnachuk, O.V., Parnachev, V.P., Holden, W., Frengstad, B., 2007. Groundwater contamination from rural pit-latrines: examples from Siberia and Kosova. Water Environ. J. 16, 147–152.

Bergé, L., 2018. Efficient Estimation of Maximum Likelihood Models With Multiple Fixed-effects: The R Package FENmlm. DEM Discussion Paper Series from Department of Economics at the University of Luxembourg.

Bergé, L., Krantz, S., McDermott, G., 2023. fixest: Fast Fixed-effects Estimations. https://github.com/lrberge/fixest/blob/master/_DOCS/FENmlm_paper.pdf. Published 2023-11-24.

Bernaisch, T., 2022. Comparing Generalised Linear Mixed-effects Models, Generalised Linear Mixed-Effects Model Trees and Random Forests: Filled and Unfilled Pauses in Varieties of English, pp. 163–193. https://doi.org/10.1017/9781108589314.007.

Bijay-Singh, Craswell, E.T., 2021. Fertilizers and nitrate pollution of surface and ground water: an increasingly pervasive global problem. SN Appl. Sci. 3.

Boumans, L.J.M., Fraters, B., 2010. A legislation induced decrease in nitrate leaching in the sandy areas of the Netherlands during the 1992–2006 period. In: Ferreira, J.A. (Ed.), RIVM Report 680717016/2010 Estimation of Net Decreases in Nitrate Concentrations. RIVM, Bilthoven, the Netherlands.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Canter, L.W., 1996. Nitrates in Groundwater. Routledge. https://doi.org/10.1201/9780203745793.

Center for International Earth Science Information Network-CIESIN-Columbia University, 2022. Global Gridded Relative Deprivation Index (GRDI), Version 1. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York.

Chambers, J.M., 1992. Linear models. In: Chapter 4 of Statistical Models in S. Wadsworth & Brooks/Cole eds.

Chambers, R., Kar, K., 2008. Handbook on Community-led Total Sanitation. IDS, Brighton.

Charulatha, G., Srinivasalu, S., Uma Maheswari, O., Venugopal, T., Giridharan, L., 2017. Evaluation of ground water quality contaminants using linear regression and artificial neural network models. Arab. J. Geosci. 10. https://doi.org/10.1007/s12517-017-2867-6.

Chavula, J., 2021. Food Is Nothing Without Hygiene. UNICEF. https://www.unicef.org/malawi/stories/food-nothing-without-hygiene. (Accessed 1 April 2024) (21 June 2021).

Chidavaenzi, M., Bradley, M., Jere, M., Nhandara, C., 2000. Pit-latrine effluent infiltration into groundwater: the Epworth case study. Schriftenr. Ver. Wasser. Boden. Lufthyg. 105, 171–177.

Chidya, R.C.G., Matamula, S., Nakoma, O., Chawinga, C.B.J., 2016. Evaluation of groundwater quality in rural-areas of northern Malawi: case of Zombwe Extension Planning Area in Mzimba. Phys. Chem. Earth 93, 55–62. https://doi.org/10.1016/j.pce.2016.03.013.

Chimphamba, J.B., Phiri, O.L., 2014. Borehole water pollution and its implication on health on the rural communities of Malawi. Malawi J. Sci. Technol. 10 (1).

Couronné, R., Probst, P., Boulesteix, A.L., 2018. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics 19. https://doi.org/10.1186/s12859-018-2264-5.

Department of Disaster Management Affairs (DoDMA) of Malawi, 2019. Malawi Floods: Situation Report 27th April 2019. United Nations Office of the Resident Coordinator in Malawi.

Dzinjalamala, G.D., Kaonga, C.C., Kumwenda, S., et al., 2024. Human health risk assessment of microbial contamina- tion and trace metals in water and soils of Chileka Township, Blantyre, Malawi. Discov. Environ. 2, 62. https://doi.org/10.1007/s44274-024-00096-4.

Dzwairo, B., Hoko, Z., Love, D., Guzha, E., 2006. Assessment of the impacts of pit-latrines on groundwater quality in rural areas: a case study from Marondera district, Zimbabwe. Phys. Chem. Earth A/B/C 31, 15–16.

Escamilla, V., Knappett, P.S.K., Mohammad Yunus, P.K.S., Emch, M., 2013. Influence of latrine proximity and type on tubewell water quality and diarrheal disease in Bangladesh. Ann. Assoc. Am. Geogr. 103, 299–308.

FAO/NASA, 2024. Malawi Cropland 2020/2021. https://data.harvestportal.org/dataset/. (Accessed 1 April 2024).

Fraser, C.M., Kalin, R.M., Kanjaye, M., Uka, Z., 2020. A methodology to identify vulnerable transboundary aquifer hotspots for multi-scale groundwater management. Water Int. 45, 865–883. https://doi.org/10.1080/02508060.2020.1832747.

Freeman, A.Y.S., Ganizani, A., Mwale, A.C., Manda, I.K., Chitete, J., Phiri, G., Stambuli, B., Chimulambe, E., Koslengar, M., Kimambo, N.R., Bita, A., Apolot, R.R., Mponda, H., Mungwira, R.G., Chapotera, G., Yur, C.T., Yatich, N.J., Totah, T., Mantchombe, F., et al., 2024. Analyses of drinking water quality during a pro-tracted cholera epidemic in Malawi – a cross-sectional study of key physicochemical and microbiological pa-rameters. J. Water Health 22 (3), 510–521.

Gilbert, M., Nicolas, G., Cinardi, G., Van Boeckel, T.P., Vanwambeke, S.O., Wint, G.R.W., Robinson, T.P., 2018. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. Sci. Data 5, 180227. https://doi.org/10.1038/sdata.2018.227.

Goldstein, B.R., de Valpine, P., 2022. Comparing N-mixture models and GLMMs for relative abundance estimation in a citizen science dataset. Sci. Rep. 12. https://doi.org/10.1038/s41598-022-16368-z.

Graham, J.P., Polizzotto, M.L., 2013. Pit latrines and their impacts on groundwater quality: a systematic review. Environ. Health Perspect. 121 (5), 521–530.

Gwenzi, W., Marumure, J., Makuvara, Z., Simbanegavi, T.T., Njomou-Ngounou, E.L., Nya, E.L., Kaetzl, K., Noubactep, C., Rzymski, P., 2023. The pit latrine paradox in low-income settings: A sanitation technology of choice or a pollution hotspot? Sci. Total Environ. 879. Preprint at doi:10.1016/j.scitotenv.2023. 163179.

Harper, C., Keith, S.M., Todd, G.D., Williams, M., Wohlers, D.W., Diamond, G.L., Coley, C., Citra, M.J., 2017. Toxicological Profile for Nitrate and Nitrite.

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. https://doi.org/10.1016/j.chemosphere.2021.133388.

Hendrickx, J., Nutricia, D., 2018. Collinearity in mixed models. In: Paper AS03. PhUSE EU Connect.

Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B.M., Gräler, B., 2018. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. https://doi.org/10.7717/peerj.5518.

Hijmans, R., 2024. raster: Geographic Data Analysis and Modeling. R Package Version 3.6-27.

Hinton, R.G., Tremblay-Lévesque, L., Macleod, C.J., Troldborg, M., Kanjaye, M., Kalin, R. M., 2024a. Progress and slippage of sanitation and hygiene targets in Malawi: is SDG6.2 achievable? J. Water Sanit. Hyg. Dev., washdev2024263 https://doi.org/10.2166/washdev.2024.263.

Hinton, R.G.K., Macleod, C.J.A., Troldborg, M., Wanangwa, G., Kanjaye, M., Mbalame, E., Mleta, P., Harawa, K., Kumwenda, S., Kalin, R.M., 2021. Factors influencing the awareness and adoption of borehole-garden permaculture in Malawi: lessons for the promotion of sustainable practices. Sustainability 13, 12196.

Hinton, R.G.K., Macleod, C.J.A., Troldborg, M., Kanjaye, M.B., Kalin, R.M., 2023. The status of sanitation in Malawi: is SDG6.2 achievable? Int. J. Environ. Res. Public Health 20.

Hinton, R.G.K., Kalin, R.M., Kanjaye, M., Mleta, P., Macleod, C.J.A., Troldborg, M., 2024b. Spatial model of groundwater contamination risks from pit-latrines in a low-income country. Water Res., 122734 https://doi.org/10.1016/j.watres.2024.122734.

Hurtt, G.C., Chini, L.P., Frolking, S., Betts, R.A., Feddema, J.J., Fischer, G.W., Fisk, J.P., Hibbard, K.A., Houghton, R.A., Janetos, A.C., Jones, C.D., Kindermann, G., Kinoshita, T., Goldewijk, K.K., Riahi, K., Shevliakova, E., Smith, S.J., Stehfest, E., Thomson, A.M., Thornton, P.E., Vuuren, D.V., Vuuren, D.V., Wang, Y., 2011. Harmonization of land-use scenarios for the period 1500–2100: 600 years of global gridded annual land-use transitions, wood harvest, and resulting secondary lands. Clim. Chang. 109, 117–161.

IHME, 2019. Global Burden of Disease Study (2019) – Processed by Our World in Data. "High Blood Pressure" [Dataset]. In: Global Burden of Disease Study. IHME.

Ishwaran, H., 2007. Variable importance in binary regression trees and forests. Electron. J. Statist. 1, 519–537. https://doi.org/10.1214/07-EJS039.

Islam, M. Sirajul, Mahmud, Zahid Hayat, Shafiqul Islam, M., Saha, Ganesh Chandra, Zahid, Anwar, Ali, Ahm Zulfiquar, Qumrul Hassan, M., Islam, Khairul, Jahan, Hasin, Hossain, Yakub, Hasan, Mahbub, Cairncross, Sandy, Carter, Richard, Luby, Stephen P., Cravioto, Alejandro, Endtz, Hubert P., Endtz, Hubert P., Faruque, Shah M., Clemens, John D., 2016. Safe distances between groundwater-based water wells and pit-latrines at different hydrogeological conditions in the Ganges Atrai floodplains of Bangladesh. J. Health Popul. Nutr. 35.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Linear Regression. An Introduction to Statistical Learning, 13. Springer, New York, NY, USA, pp. 32–58.

Jung, H., Koh, D.-C., Kim, Y.S., Jeen, S.-W., Lee, J., 2020. Stable isotopes of water and nitrate for the identification of groundwater flowpaths: a review. Water 12, 138. https://doi.org/10.3390/w12010138.

Kalin, R.M., Mwanamveka, J., Coulson, A.B., Robertson, D.J., Clark, H., Rathjen, J.P., Rivett, M.O., 2019. Stranded assets as a key concept to guide investment strategies for sustainable development goal 6. Water 11, 702.

Kalin, R.M., Mleta, P., Addison, M.J., Banda, L.C., Butao, Z., Nkhata, M., Rivett, M.O., Mlomba, P., Phiri, O., Mambulu, J., Phiri, O.C., Kambuku, D.D., Manda, J., Gwedeza, A., Hinton, R., 2022a. Hydrogeology and Groundwater Quality Atlas of Malawi, Bulletin. Ministry of Water and Sanitation, Government of Malawi. ISBN 978-1-915509-00-0 151pp.

Kalin, R.M., Mleta, P., Addison, M.J., Banda, L.C., Butao, Z., Nkhata, M., Rivett, M.O., Mlomba, P., Phiri, O., Mambulu, J., Phiri, O.C., Kambuku, D.D., Manda, J., Gwedeza, A., Hinton, R., 2022b. Hydrogeology and Groundwater Quality Atlas of Malawi, Linthipe River Catchment, Water Resource Area 4. Ministry of Water and Sanitation, Government of Malawi. ISBN 978-1-915509-05-5 110pp.

Karunanidhi, D., Subramani, T., Roy, P.D., Li, H., 2021. Impact of groundwater contamination on human health. Environ. Geochem. Health. https://doi.org/10.1007/s10653-021-00824-2.

Kayembe, J.M., Thevenon, F., Laffite, A., Sivalingam, P., Ngelinkoto, P., Mulaji, C.K., Otamonga, J.P., Mubedi, J.I., Poté, J., 2018. High levels of faecal contamination in drinking groundwater and recreational water due to poor sanitation, in the sub-rural neighbourhoods of Kinshasa, Democratic Republic of the Congo. Int. J. Hyg. Environ. Health 221, 400–408. https://doi.org/10.1016/j.ijheh.2018.01.003.

Kelly, L., Bertram, D., Kalin, R.M., Ngongondo, C., Sibande, H., 2020. A national scale assessment of temporal variations in groundwater discharge to rivers: Malawi. Am. J. Water Sci. Eng. 6 (1), 39–49. https://doi.org/10.11648/j.ajwse.20200601.15.

Kendall, C., Elliott, E.M., Wankel, S.D., 2007. Tracing anthropogenic inputs of nitrogen to ecosystems. In: Michener, R., Lajtha, K. (Eds.), Stable Isotopes in Ecology and Environmental Science. https://doi.org/10.1002/9780470691854.ch12.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28. https://doi.org/10.18637/jss.v028.i05.

Kundzewicz, Z.W., Döll, P., 2009. Will groundwater ease freshwater stress under climate change? Hydrol. Sci. J. 54, 665–675. https://doi.org/10.1623/hysj.54.4.665.

Li, P., Karunanidhi, D., Subramani, T., Srinivasamoorthy, K., 2021. Sources and consequences of groundwater contamination. Arch. Environ. Contam. Toxicol. https://doi.org/10.1007/s00244-020-00805-z.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News 2, 18–22.

Liu, C., Bartlet-Hunt, S., Li, Y., 2024. Precipitation, temperature, and landcovers drive spatiotemporal variability of groundwater nitrate concentration across the Continental United States. Sci. Total Environ. 945 (2024), 174040. ISSN 0048-9697. https://doi.org/10.1016/j.scitotenv.2024.174040.

Liu, X., 2016. Generalized linear mixed models on nonlinear longitudinal data. In: Methods and Applications of Longitudinal Data Analysi, pp. 243–279. https://doi.org/10.1016/b978-0-12-801342-7.00008-3.

Louppe, G., 2014. Understanding Random Forests: From Theory to Practice. arXiv: Machine Learning.

Malawi Bureau of Standards (MBS), 2017. Catalogue of Malawi standards. In: Malawi Bureau of Standards. Moirs Road P.O Box 946 Blantyre Malawi.

Mas-Pla, J., Menció, A., 2019. Groundwater nitrate pollution and climate change: learnings from a water balance-based analysis of several aquifers in a western Mediterranean region (Catalonia). Environ. Sci. Pollut. Res. 26, 2184–2202. https://doi.org/10.1007/s11356-018-1859-8.

Mayer, M., Stando, A., 2024. SHAP Visualizations. URL. https://github.com/ModelOriented/shapviz.

Mayer, M., Watson, D., Biecek, P., 2023. Kernel SHAP. URL. https://github.com/ModelOriented/kernelshap.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. Front. Econ. 105–142.

McFadden, D., 1977. Quantitative methods for analyzing travel behaviour of individuals: some recent developments. In: Cowles Foundation Discussion Papers, 707.

Minet, E.P., Goodhue, R., Meier-Augenstein, W., Kalin, R.M., Fenton, O.K., Richards, K. G., Coxon, C.E., 2017. Combining stable isotopes with contamination indicators: A method for improved investigation of nitrate sources and dynamics in aquifers with mixed nitrogen inputs. Water Res. 124, 85–96. ISSN 0043-1354. https://doi.org/10.1016/j.watres.2017.07.041.

Ministry of Agriculture, Irrigation and Water Development (MAIWD), 2018. National Agricultural investment plan. Prioritised and Coordinated Agricultural Transformation Plan for Malawi: FY 2017/18-2022/23.

Mkandawire, T., 2008. Quality of groundwater from shallow wells of selected villages in Blantyre District, Malawi. Phys. Chem. Earth 33, 807–811.

Morin, K.A., Hutt, N.M., 2009. Mine-water Leaching of Nitrogen Species From Explosive Residues.

Morrissy, Justin G., Currell, Matthew J., Reichman, Suzie M., Surapaneni, Aravind, Megharaj, Mallavarapu, Crosbie, Nicholas D., Hirth, Daniel, Aquilina, Simon, Rajendram, William, Ball, Andrew S., 2021. Nitrogen contamination and bioremediation in groundwater and the environment: a review. Earth Sci. Rev. 222, 103816. https://doi.org/10.1016/j.earscirev.2021.103816.

Muschelli, J., Betz, J., Varadhan, R., 2014. Chapter 7- binomial regression in R. In: Rao, M.B., Rao, C.R. (Eds.), Handbook of Statistics, pp. 257–308. https://doi.org/10.1016/B978-0-444-63431-3.00007-3.

Mussa, C., Kamoto, J.F., 2023. Groundwater quality assessment in urban areas of Malawi: a case of area 25 in Lilongwe. J. Environ. Public Health 2023, ID6974966. https://doi.org/10.1155/2023/6974966.

Nakagawa, K., Amano, H., Persson, M., Berndtsson, R., 2021. Spatiotemporal variation of nitrate concentrations in soil and groundwater of an intensely polluted agricultural area. Sci. Rep. 11, 2598. https://doi.org/10.1038/s41598-021-82188-2.

Nath, B., Chowdhury, R., Ni-Meister, W., Mahanta, C., 2022. Predicting the distribution of arsenic in ground- water by a geospatial machine learning technique in the two most affected districts of Assam, India: the public health implications. Geohealth 6. https://doi.org/10.1029/2021GH000585.

National Planning Commission (NPC), 2021. Malawi 2063: Malawi's Vision an Inclusively Wealthy and Self-reliant Nation. National Planning Commission (NPC), Government of Malawi, Lilongwe, Malawi.

National Statistical Office (NSO), ICF, 2017. Malawi Demographic and Health Survey 2015–16. Key Indicators Report. The DHS Program. ICF International.

National Statistical Office (NSO), Malawi, 2021. Malawi Multiple Indicator Cluster Survey 2019–20 (MICS) Survey Findings Report. National Statistical Office, Malawi.

Ndoziya, A.T., Hoko, Z., Gumindoga, W., 2019. Assessment of the impact of pit-latrines on groundwater contamination in Hopley Settlement, Harare, Zimbabwe. J. Water Sanit. Hyg. Dev. 9, 464–476.

Nielsen, A.M., Garcia, L.A.T., Silva, K.J.S., Sabogal-Paz, L.P., Hincapié, M.M., Montoya, L.J., Galeano, L., Galdos-Balzategui, A., Reygadas, F., Herrera, C., Golden, S., Byrne, J.A., Fernández-Ibáñez, P., 2022. Chlorination for low-cost household water disinfection – a critical review and status in three Latin American countries. Int. J. Hyg. Environ. Health 244.

Nikolenko, O., Jurado, A., Borges, A.V., Knöller, K., Brouyère, S., 2018. Isotopic composition of nitrogen species in groundwater under agricultural areas: a review. Sci. Total Environ. 621, 1415–1432, 2018. ISSN 0048-9697. https://doi.org/10.1016/j.scitotenv.2017.10.086.

Nkwanda, I.S., Feyisa, G.L., Zewge, F., Makwinja, R., 2021. Impact of land-use/land-cover dynamics on water quality in the Upper Lilongwe River basin, Malawi. Int. J. Energy Water Resour. 5, 193–204. https://doi.org/10.1007/s42108-021-00125-5.

Nolan, B.T., Hitt, K.J., 2006. Vulnerability of shallow groundwater and drinking-water wells to nitrate in the United States. Environ. Sci. Technol. 40, 7834–7840. https://doi.org/10.1021/es060911u.

Nyenje, P.M., Foppen, J.W., Uhlenbrook, S., Uhlenbrook, S., Kulabako, R., Muwanga, A., 2010. Eutrophication and nutrient release in urban areas of sub-Saharan Africa–a review. Sci. Total Environ. 408 (3), 447–455.

Okullo, J.O., Moturi, W.N., Ogendi, G.M., 2017. Open defaecation and its effects on the bacteriological quality of drinking water sources in Isiolo County, Kenya. Environ. Health Insights 11. https://doi.org/10.1177/117 8630217735539.

Ouedraogo, I., Defourny, P., Vanclooster, M., 2019. Application of random forest regression and comparison of its performance to multiple linear regression in modeling groundwater nitrate concentration at the African continent scale. Hydrogeol. J. 27, 1081–1098. https://doi.org/10.1007/s10040-018-1900-5.

Potter, P., Ramankutty, N., Bennett, E.M., Donner, S.D., 2010. Characterizing the spatial patterns of global fertilizer application and manure production. Earth Interact. 14, 1–22.

Potter, P., N.R.E.M.B, S.D.D, 2012. Global Fertilizer and Manure, Version 1: Nitrogen in Manure Production. NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York. https://doi.org/10.7927/H4KH0K81. Accessed 19/01/2024.

Pritchard, M., Mkandawire, T., O'Neill, J.G., 2007. Biological, chemical and physical drinking water quality from shallow wells in Malawi: case study of Blantyre, Chiradzulu and Mulanje. Phys. Chem. Earth A/B/C 32, 1167–1177.

Pritchard, M., Mkandawire, T., O'Neill, J.G., 2008. Assessment of groundwater quality in shallow wells within the southern districts of Malawi. Phys. Chem. Earth A/B/C 33 (8–13), 812–823.

Puckett, L.J., Tesoriero, A.J., Dubrovsky, N.M., 2011. Nitrogen contamination of surficial aquifers–a growing legacy. Environ. Sci. Technol. 45 (3), 839–844. https://doi.org/10.1021/es1038358.

Pullanikkatil, D., Palamuleni, L.G., Ruhiiga, T.M., 2015. Impact of land use on water quality in the Likangala catchment, southern Malawi. Afr. J. Aquat. Sci. 40, 277–286. https://doi.org/10.2989/16085914.2015.1077777.

QGIS Development Team, 2024. QGIS geographic information system. In: Open Source Geospatial Foundation Project. http://qgis.osgeo.org.

R Core Team, 2023. R: A Language and Environment for Statistical Computing.

Rabe-Hesketh, S., Skrondal, A., 2008. Generalized Linear Mixed Models. Longitudinal Data Analysis. Chapman and Hall/CRC, 9780429142673.

Rahman, A., Mondal, N.C., Tiwari, K.K., 2021. Anthropogenic nitrate in groundwater and its health risks in the view of background concentration in a semi-arid area of Rajasthan, India. Sci. Rep. 11, 9279.

RCMRD, 2015a. Malawi National Anthropogenic Biomes. Sept. 28, 2015.

RCMRD, 2015b. Malawi National Precipitation Trend. Sept. 24, 2015.

Rivett, M.O., Tremblay-Levesque, L.-C., Carter, R., Thetard, R.C.H., Tengatenga, M., Phoya, A., Mbalame, E., Mchilikizo, E., Kumwenda, S., Mleta, P., Addison, M.J., Kalin, R.M., 2022. Acute health risks to community hand-pumped groundwater supplies following Cyclone Idai flooding. Sci. Total Environ. 806 (Part 2).

Rokach, L., Maimon, O., 2015. Data mining with decision trees: theory and applications. In: Series in Machine Perception Artificial Intelligence, 81. ISBN 978-9814590075.

Sajidu, S.M.I., Masamba, W.R.L., Henry, E.M.T., Kuyeli, S.M., 2007. Water quality assessment in streams and wastewater treatment plants of Blantyre, Malawi. Phys. Chem. Earth 32, 1391–1398. https://doi.org/10.1016/j.pce.2007.07.045.

Scottish Government, 2019. Contribution to International Development Report 2018–2019. External Affairs Directorate. ISBN 9781839601644.

Sridhar, D., Parimalarenganayaki, S., 2024. A comprehensive review on groundwater contamination due to sewer leakage: sources, detection techniques, health impacts, mitigation methods. Water Air Soil Pollut. https://doi.org/10.1007/s11270-023-06852-1.

Templeton, M.R., Hammoud, A.S., Butler, A.P., Braun, L., Foucher, J., Grossmann, J., Boukari, M., Faye, S., Jourda, J.P., 2015. Nitrate pollution of groundwater by pit latrines in developing countries. AIMS Environ. Sci. 2 (2), 302–313.

Tillett, T., 2013. Pit-latrines and groundwater contamination: negative impacts of a popular sanitation method. Environ. Health Perspect. 121, 5.

Truslove, J.P., V. M. Miller, A., Mannix, N., Nhlema, M., Rivett, M.O., Coulson, A.B., Mleta, P., Kalin, R.M., 2019. Understanding the functionality and burden on decentralised rural water supply: influence of millennium development goal 7c coverage targets. Water 11, 494. https://doi.org/10.3390/w11030494.

Truslove, J.P., Coulson, A.B., Mbalame, E., Kalin, R.M., 2020. Barriers to hand-pump serviceability in Malawi: life-cycle costing for sustainable service delivery. Environ. Sci. Water Res. Technol. 6, 2138–2152. https://doi.org/10.1039/D0EW00283F.

UNICEF, 1995. Monitoring Progress Toward the Goals of the World Summit for Children a Practical Handbook for Multiple-indicator Surveys. Planning Office Evaluation and Research Office Programme Division.

UNICEF, 2016. Manual for Water Quality Testing.

UNICEF, 2018. UNICEF's game plan to end open defecation. www.washdata.org.

UNICEF, WHO, 2024. Malawi household safe drinking water. https://washdata.org/data/household#!/table?geo0=country&geo1=MWI. (Accessed 25 January 2024).

United Nations, D. of E. and S.A.P.D, 2024. Malawi Population Projection. https://population.un.org/wpp/. (Accessed 15 January 2024).

von Hellens, A., 2013. Groundwater quality of Malawi– fluoride and nitrate of the Zomba-Phalombe plain. In: Degree project in Biology Agriculture Programme-Soil and Plant Sciences Examensarbeten, Institutionen för mark och miljö, SLU Uppsala 2013 2013:10. Accessed. https://stud.epsilon.slu.se/5651/1/von_hellens_a_130611.pdf.

WHO, 2017. Guidelines for Drinking-Water Quality: Fourth Edition Incorporating the First Addendum. World Health Organization, Geneva, 2017. PMID: 28759192.

Wick, K., Heumesser, C., Schmid, E., 2012. Groundwater nitrate contamination: factors and indicators. J. Environ. Manag. 111, 178–186. https://doi.org/10.1016/j.jenvman.2012.06.030.

Wilkinson, G.N., Rogers, C.E., 1973. Symbolic descriptions of factorial models for analysis of variance. Appl. Stat. 22, 392–399.

World Bank, 2024. World Bank Population Estimate Malawi. https://data.worldbank.org/country/MW. (Accessed 30 March 2024).

World Health Organization (WHO), United Nations Children's Fund (UNICEF), 2017. Safely Managed Drinking Water: Thematic Report on Drinking Water 2017. World Health Organization. https://iris.who.int/handle/10665/325897 (License: CC BY-NC-SA 3.0 IGO).

Worldpop, 2024. www.worldpop.org. (Accessed 30 April 2024).

Wright, J., Gundry, S., Conroy, R., 2004. Household drinking water in developing countries: a systematic review of microbiological contamination between source and point-of-use. Trop. Med. Int. Health 1, 106–117. https://doi.org/10.1046/j.1365-3156.2003.01160.x.

Wright, J.A., Cronin, A., Okotto-Okotto, J., 2013. A spatial analysis of pit-latrine density and groundwater source contamination. Environ. Monit. Assess. 185, 4261–4272.

Zhu, H., Liang, F., Gu, M., Peterson, B.S., 2007. 18- stochastic approximation algorithms for estimation of spatial mixed models. In: Lee, S.-Y. (Ed.), Handbook of Latent Variable and Related Models. North-Holland, Amsterdam, pp. 399–421. https://doi.org/10.1016/B978-044452044-9/50021-5.